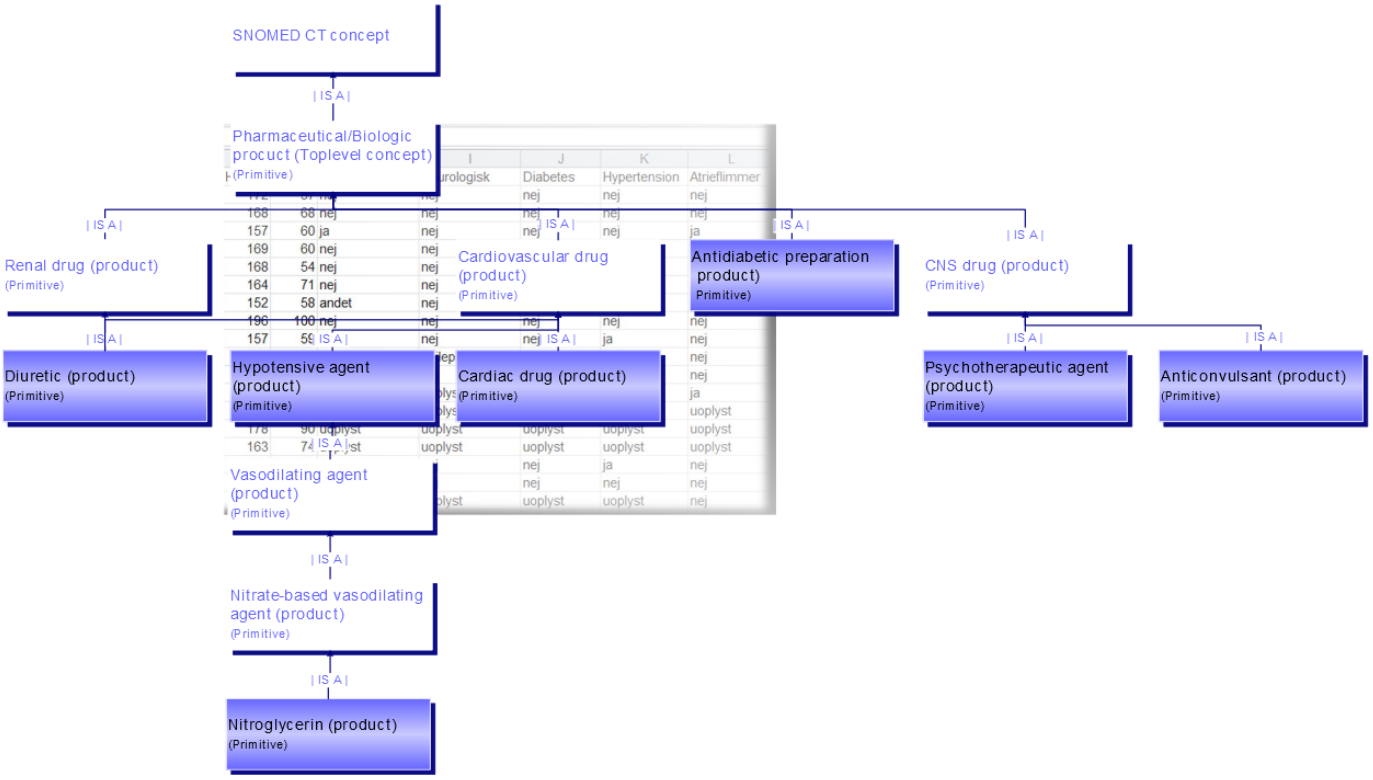
Exploring outcomes of research data structured by SNOMED CT

- An exemplification based on mapping a syncope research dataset



Master Thesis

Department of Health Science and Technology |
Biomedical Engineering and Informatics |
Camilla Lærke Steenvinkel | Jan. 2014 |
Article|

AALBORG UNIVERSITY
DENMARK

Exploring outcomes of research data structured by SNOMED CT

- An exemplification based on mapping a syncope research dataset

Camilla Steenvinkel[1]*

Abstract

Background: When mapping clinical information to SNOMED CT it is important that mapping is kept consistent. But selection of SNOMED CT concepts is ambiguous. Therefore, mapping guidelines are necessary to ensure consistent mapping. Only limited instructions of how to map clinical information to SNOMED CT are available. [1] have developed a mapping guideline for mapping EHR-template terms to SNOMED CT. No research studies of how research data should be mapped to SNOMED CT were found by the literature review conducted in this project.
The objective of this project was to investigate the applicability of [1]s' mapping guideline for mapping a research dataset and to investigate how this mapping guideline should be adapted to facilitate mapping of research data. Further, it was investigated what SNOMED CT may add to a research dataset which is mapped by use of this mapping guideline.
Method: Investigation was conducted by an exemplification, where a research dataset of 941 syncope patients was mapped to SNOMED CT. The mapping process involved the 3 steps: Grouping of the DEs of the dataset, mapping each group, and refinements of the mapping. Mapping was conducted by an iterative mapping process and refinements were conducted until the mapped DEs fulfilled a set of quality criteria. These quality criteria were used to evaluate the quality of mapping, thus ensure consistency. Selection of SNOMED CT concepts was conducted according to [1]s' mapping guideline. For areas which this guideline did not cover mapping was conducted in 3 steps; 1) candidate concepts were identified. 2) An overview of the possibilities and limitations of each candidate was provided by drawing the defining- and qualifying relationships of each candidate concept. 3) The best candidate was selected.
Results: The DEs of the dataset was divided into 6 groups. Of these, two groups were selected ("'diagnoses"' and "'medication"') and mapped to SNOMED CT. It was possible to map the DEs of each group to subtype descendants of the OE of each group, respectively. Thus, each group created a subset cluster where the OE was the LCP of the DEs of the group.
Since it was not possible to find appropriate SNOMED CT concepts to represent the contextual data values of the dataset ("'Yes"'/"'No"'/"'Unspecified"') these where not mapped to SNOMED CT. 1 DE was not mapped, since it was not possible to interpret its semantic meaning.
Conclusion: SNOMED CT provides representation of research data with optional level of granularity. Further, SNOMED CT adds more details to the dataset. SNOMED CT based research data is one step towards semantic interoperability and efficient data extraction from EHR-systems, thus one step towards efficient, high-quality translational research and improved outcome of the clinical care process.

Abbreviations:

- CIS - Clinical Information System.
- CMV - Controlled Medical Vocabulary.
- DE - Data Element.
- DNPR - the Danish National Patient Register. In Danish; Landpatientregisteret (LPR).
- ICD-10 - The International Classification of Diseases the 10th revision.
- LCP - Least Common Parent.
- SNOMED CT - the Systematized Nomenclature of Medicine Clinical Terms.
- OE - Organising Element.

Keywords
SNOMED CT — mapping guideline — mapping instructions — rule development — encoding — mapping method — terminology mapping — semantic interoperability — information retrieval

[1] *Department of Health Science and Technology, Aalborg University, Aalborg, Denmark*
***Corresponding author**: csteen07@student.aau.dk

Along with this article a set of worksheets are provided to support the reading. The worksheets also includes a nomenclature of terms and abbreviations. Circular references to the worksheets are added where needed.

Contents

1. Introduction

1.1 Research and standardisation of research data

Research data is generally used for research studies, to discover new knowledge and more details about a topic.

Clinical research may change between setting up hypotheses and conformation/disconfirmation of these hypotheses. The hypotheses may be set up based on statistical analyses. The hypotheses may be confirmed/disconfirmed by clinical trials, which purpose is to study further details of clinical data and findings that do not properly fit the expected results or hypotheses.

Statistical analyses require uniquely assigned Data Elements (DE) with a high degree of data completeness. Studying further details may require data with a high level of granularity and correlations between DEs to reflect the details of the real world. Classification systems provide these uniquely assigned DEs appropriate for statistical analyses, while concept systems provide more details to reflect the true details of the real world [2]. A complete clarification of the advantages and weaknesses of classification systems and concept systems, respectively is provided in chapter 1 of the appendix report. Further, this chapter clarifies how these types of Controlled Medical Vocabularies (CMV) should be implemented and interact in a Clinical Information Systems (CIS).

This project concerns mapping a research dataset of 941 syncope patients to SNOMED CT to investigate how research data should be mapped to SNOMED CT and what SNOMED CT may add to research data. The syncope patients had underwent a diagnosing commitment in order to determine the underlying patophysiological cause of their syncope episodes.

A syncope episode is characterised by a Transient Loss of Consciousness (T-LOC) associated with decreased tonus of skeletal muscles and complete spontaneous recovery. Remission is rapid and spontaneous, but in some cases fatigue present in the post-recovery period. A syncope episode is caused by global cerebral hypoperfusion. Syncope are divided into three main groups according to the underlying pathophysiological cause; Cardiovascular syncope, syncope secondary to orthostatic hypotension, and reflex syncope. [3]

Evaluation of syncope patients is important to determine the cause of the syncope episodes in order to address an effective treatment and to identify if any underlying disease (e.g. cardiovascular disorders) is the reason of the syncope episodes. [3]

The principle of diagnosing syncope is to provoke a syncope episode (or the characteristic changes in pulse and blood pressure dependent on the method used) while measuring pulse, standard ECG, and blood pressure. This is done by a head-up tilt test. Provocation of a syncope episode may include carotid massage and/or medication. [4]

A head-up tilt test provokes a syncope episodes in 25% of the patients who suffer from syncope episodes. 50 % of all patients who undergo a diagnosing commitment remains undiagnosed. Some medication (isoprenaline and nitro-glycerine) increases the change of provocation of a syncope episode. Reproduction of the test results from the head-up tilt test is difficult, especially the positive test results are difficult to reproduce.[4]

Since diagnosing relies upon measurement of pulse, ECG, and blood pressure during a syncope episode, diagnosing by use of the head-up-tilt test is difficult. Further, the problems with reproduction makes the test method less reliable. [4]

Standardisation of research data

CIS used for routine clinical use and research projects, respectively, are developed independently of each other. Therefore, CIS used for routine clinical use and CIS used for research projects do not have a common data representation and a common data dictionary which causes that data from each research study is represented independently by its own proprietary DEs. [5] Use of free text and several terminologies which have partial and overlapping domain coverage hinders reaching semantic interoperability, thus hinders enabling secondary use of clinical data [6].

To facilitate unambiguous and standardised representation- and interpretation of research data standardisation is necessary. Further, standardisation is expected to enable data aggregation of research data. [5] Standardised data is important in clinical practice as well as in clinical research to reach semantic interoperability, thus facilitate reuse of clinical data for clinical research. Providing reuse of standardised clinical data will increase the quality and speed of clinical- and translational research, thus improve clinical prac-

tice and public health. [5]

1.2 SNOMED CT

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a scientifically validated terminology. SNOMED CT is a comprehensive and compositional reference terminology characterised by a poly-axial hierarchy and defining characteristics between concepts which are inherited from the Concept Model of SNOMED CT. The absolute position of a SNOMED CT concept defines its semantic meaning. Thus, SNOMED CT provides explicit semantic meaning of information and multilingualism. [7] Chapter 2 of the appendix report explains the basics about SNOMED CT concepts, the subtype hierarchies, defining characteristics, post-coordination, and the Compositional Model of SNOMED CT. Further, some problematic issues of SNOMED CT are clarified.

What does SNOMED CT provide?

SNOMED CT provides standardisation of clinical information, thus provides unambiguous representation of clinical terms [8, 9]. SNOMED CT provides high expressiveness and details compared to other terminologies and classification systems. It is possible to cross map SNOMED CT-based data to ICD, e.g. to provide statistics. SNOMED CT is excellent for data retrieval purposes. [9] E.g. SNOMED CT can provide retrieval of patients with chronic disorders, by query for all disorders which have the defining characteristic |Clinical cause (Linkage concept)| |Chronic (qualifier value)|. This is not possible by ICD unless each disease-code is "'handpicked"'. If a new chronic disorder is added to the ICD-hierarchy or if changes of the disease-code are applied (e.g. due to a new version of ICD) the list of "'handpicked"' disease-codes must be configured. In SNOMED CT it does not matter if a new chronic disorder is added to the terminology as long as the new SNOMED CT concept also has the defining characteristic |Clinical cause (Linkage concept)| |Chronic (qualifier value)|. Further, information retrieval in an EHR by use of SNOMED CT concepts have shown 25% better precision than keyword search [6].

Despite of inconsistency problems and lack of SNOMED CT concepts, SNOMED CT provides the best content coverage compared to other terminologies [9]. High content coverage for DEs commonly used in clinical practice is experienced [5]. From a review of practical implementations of SNOMED CT it was found that 19%-90% of the mapped clinical terms were mapped to exact matches [9]. This review did not clarify any reason of this high variation, but three reasons may affect the rate of exact matches; First, variation of the definition of exact matches exists. Second, some clinical domains and some subtype hierarchies within SNOMED CT have a higher maturity level than others. Therefore, the content coverage depends on the clinical domain to be mapped. Third, DEs could be inappropriately defined. Two mapping approaches showed that 1% of clinical terms which could not be mapped to SNOMED CT [8, 10]. In another mapping approach 12% of the EHR-template DEs could not be mapped to SNOMED CT, but it was discussed that the not-mappable DEs were defined inappropriately [1].

SNOMED CT provides post-coordination which is the combination of two SNOMED CT concepts, linked by a SNOMED CT attribute concept. [7] Additional details of pre- and post-coordination are clarified in chapter 2, section 2.2 of the appendix report. This section also clarifies some of the shortcomings of post-coordination. Post-coordination is an important feature because it allows to express detailed and special instances of a concept and to express more detailed expressions without expanding the number of pre-coordinated concepts within a terminology. [5] Therefore, post-coordination provides higher expressiveness to the terminology.

SNOMED CT also includes detailed expressions within a single SNOMED CT concept and these concepts are denoted as pre-coordinated concepts. Pre- and post-coordination provide similar and equal expressions defined by different concepts and attributes. [11] As an example:

- The pre-coordinated SNOMED CT concept |Hypotensive syncope (Clinical finding)| may be expressed by the post-coordinated expression |Syncope (Clinical finding)| |DUE TO (Linkage concept) | |Low blood pressure (Clinical finding)|.

These two expressions are semantically equal, even though they are expressed with different SNOMED CT concepts. Pre- and post-coordination provides SNOMED CT with compositional characteristics [1].

What are the current problems of SNOMED CT?

Several studies have found problems with content coverage and inconsistency of SNOMED CT [10, 9, 1, 8, 5]. The content coverage problems are lack of SNOMED CT concepts and defining relationships. It is experienced that SNOMED CT does not provide representation of medication and ingredients very well. Inconsistency problems covers inconsistent intermediate subtype relationships and missing subtype relationships. [9] Inconsistent naming of SNOMED CT concepts is also found. E.g. the two SNOMED CT concepts |Left popliteal artery structure (Body structure)| and |Structure of right popliteal artery (Body structure)| are named differently. [10] Further, SNOMED CT does not provide adequate representation of negation [9].

Mapping clinical datasets to SNOMED CT may often require post-coordination. [9] found that in 40% of the investigated mapping approaches post-coordination was required. From three practical mapping approaches post-coordination was used for 23-80% of the data items [1, 10, 5]. Post-coordination is complex for three reasons. First, it is difficult to design a simple UI where users can understand how to select appropriate refinements. Second, many issues of post-coordination are not investigated yet, e.g. is not investigated how post-coordinated expressions can be mapped to classification systems as ICD-10. Third, post-coordination can lead to inconsistency and redundancy. [9] E.g. from the example mentioned above it is only possible to automatically compute that |Syncope (Clinical finding)| |DUE TO (Linkage concept)| |Low blood pressure (Clinical finding)| is semantically equal to |Hypotensive syncope (Clinical finding)| if this pre-coordinated SNOMED CT concept is fully defined. If |Hypotensive syncope (Clinical finding)| was primitive there would be a risk of redundant information of pre- and post-coordinated expressions within a dataset which could not be interpreted as semantically equal by computer processing.

1.3 Mapping research data to SNOMED CT

SNOMED CT should be implemented in conjunction with an information model. This conjunction should provide representation of dates, numeric numbers, and other information which should not be represented within a terminology. [9]

Challenges of mapping

The multi-axial hierarchy of SNOMED CT is large and complex. Selection of SNOMED CT concepts during mapping is ambiguous for three reasons. First, to map a clinical term correctly it is important to understand its intended semantic meaning [8]. Second, intercoder variation of interpretation of SNOMED CT concepts is experienced [10]. Third, when mapping a DE to SNOMED CT there may be more than a single SNOMED CT concept which seems to be obvious to select to represent that DE, but not all options provide the same opportunities and limitations for data retrieval. Especially when post-coordination is required this is true. Therefore, intercoder variation appears if no mapping rules are used [12].

Manual mapping of clinical information is tedious and time-consuming [8]. For automatic mapping methods cleaning of data items is often necessary. Data cleaning is also a time-consuming process [8].

In a cross-mapping approach it was showed that intercoder reliability could be significantly improved by mapping rules. [12]

Several research approaches have found consistency problems of mapping clinical terms to SNOMED CT, but none of these studies have defined what consistent mapping is [12, 13, 1, 10]. *In this project consistent mapping is defined as similar mapping of equal concepts with minimum intercoder- and intra coder variation. Equal concepts may be considered as semantically equal concepts (e.g. diagnoses, procedures etc.) or equal data types (list items, free text, measurements with numerical values, dates etc.).* The semantically grouping defines which subtype hierarchy concepts should be mapped to while the data types defines how the DE should be represented by SNOMED CT concepts, e.g. how post-coordination should be created.

Consistent mapping is necessary to overcome the inconsistency- and redundancy problems of SNOMED CT. Further, consistent mapping is necessary to provide data retrieval. [1]

Therefore, when mapping clinical information it is critical to have terminology experts who have knowledge about terminologies, SNOMED CT, and health professional knowledge. These experts are able to contribute with their knowledge for data items and concepts which semantic meaning are not obvious by their description.

[10] A combined knowledge provides the ability to see the big picture. E.g. when defining groups it is beneficial to consider how to define groups which could be mapped to a subset cluster of SNOMED CT concepts. (In this project a subset cluster is defined as a set of SNOMED CT concepts which are inherited from the same Least Common Parent concept (LCP). The LCP is the nearest supertype ancestor of a set of SNOMED CT concepts. The LCP is included in the subset cluster. A subset cluster can be obtained as a view of SNOMED CT concepts.) In that way a combined knowledge makes mapping easier since all concepts within a group should be mapped similarly and the group name helps identifying a LCP SNOMED CT concept of a subset cluster in SNOMED CT.

For both manual and automatic mapping careful expert review of the mapped terms is necessary to ensure appropriate selection of SNOMED CT concepts [8, 1, 10].

Mapping guideline research

To ensure consistent mapping and overcome the limitations in SNOMED CT there is a need for consensus about the use of SNOMED CT [1] by development of clear/detailed mapping instructions. Access to detailed mapping instructions will also improve the applicability of SNOMED CT.

Only limited detailed instructions- and examples of how SNOMED CT should be implemented into a CIS exist [9, 6]. The available tutorials and educational workshops mainly focus on general aspects of SNOMED CT - Not on how to map and implement SNOMED CT. Learning general aspects of SNOMED CT is necessary for people who are completely new to SNOMED CT, but practical examples are necessary to be able to implement SNOMED CT in clinical (research) settings. [8]

From one study it was found that the official guidelines of IHSTDO were inadequate to ensure consistent mapping of EHR-template content to SNOMED CT. It was suggested to extend the official guideline of the International Health Terminology Standards Development Organisation (IHTSDO) to ensure standardised and more comprehensive mapping. [1] In a survey of SNOMED CT implementations it was found that implementers would like more instructions from IHTSDO of how to map clinical information to SNOMED CT. There is a need of clear methodologies of how to develop appropriate subset clusters. Especially for mapping of broad domains which have shown to be difficult to map. The DE "'Reason for admission"' is an example of a broad domain since the data values of such a DE could be nearly everything and within different domains. [9]

Several studies have mentioned possible benefits of SNOMED CT, but no studies have investigated the value of SNOMED CT in terms of improved outcome and no studies have investigated the clinical value of SNOMED CT in clinical practice [14, 9, 6]. SNOMED CT is still used for low level implementation approaches. Focus has been on technical issues as data entry and data retrieval. This shows that use and implementation of SNOMED CT is still in its early phase of implementation and many issues are not investigated yet. Therefore the high level benefits as decision support, semantic interoperability, improved patient care etc. are not obtained yet. [9, 6]

From a literature review conducted during this project it was found that only limited research within mapping guideline research exists. Most research approaches focused on mapping clinical information to SNOMED CT by use of semi-automatic lexical matching techniques [5, 8, 10]. Only limited considerations of how mapping was kept consistent were made within these studies. For the few guidelines used in these studies the authors have spend time to discuss how their data should be mapped to SNOMED CT in order to reach consensus and develop their own "'internal mapping guidelines"' which concerned some sets of rules for which subtype hierarchy to select from. However, it is important to consider semantical and structural issues when mapping to SNOMED CT. Only a single research approach about mapping guideline research to facilitate consistent mapping was found ([1]). This study considered consistent mapping of EHR-template terms into SNOMED CT. EHR-template terms are naturally grouped by Organising Elements (OE) and the mapping guideline used these groups to set-up at set of rules of which top-level hierarchies to select from. Further, the mapping guideline included at set of rules of how to represent different data types by SNOMED CT concepts. This mapping guideline only considered diagnoses and procedures.

1.4 Summary of what mapping guidelines should include

From the introduction, in this section, it is found that current literature supports the need of development and research of mapping guidelines to ensure consistent mapping. These mapping guidelines should include the following elements:

1. Guidelines of how non-grouped clinical data should be grouped prior to mapping.
2. Detailed explanations- and examples of when to select a specific subtype hierarchy and when not to select it.
3. Guidelines of how different data types should be represented by SNOMED CT concepts. These guidelines should also specify which parts of a DE that should be represented by SNOMED CT and which parts should be represented by an information model.
4. Detailed mapping instructions should both include a set of rules (mapping guidelines) and a mapping procedure.
5. Instructions and examples of special complex cases.
6. Guidelines for when to- and how to redefine DEs.

These guidelines should come along with detailed- and illustrative examples.

1.5 Objective of the project

The aim of this project is to contribute to the development of detailed mapping instructions and examples of how research data should be mapped to SNOMED CT, thus contribute to improve consistency of mapping and the applicability of SNOMED CT.

Therefore, this project concerns the question: *How is it possible to use an existing mapping guideline for EHR-templates to map a research dataset? - And what adaptions should be applied to that mapping guideline to facilitate mapping of research data? Further, what may SNOMED CT add to research data by use of this mapping guideline to map a research dataset to SNOMED CT?*

This project focused on the first four elements of the list mentioned in section 1.4. These four elements were selected since these items were considered to be essential to ensure consistent mapping the most important, while the two remaining elements were supportive guidelines for more advanced mapping topics.

It was investigated if [1]s' mapping guideline could be used to map research data and it was investigated what extensions and/or modifications of their mapping guideline that were necessary to implement to be able to map research data. First, the mapping guideline was extended/modified. Second, a syncope research dataset was mapped to SNOMED CT by use of this modified mapping guideline. The exemplification served as an evaluation of the applicability of the mapping guideline and provided an example of how a research dataset should be mapped to SNOMED CT. Further, the exemplification showed the possibilities and limitations of a SNOMED CT based research dataset which is mapped by use of this mapping guideline.

2. Method

An overview of the methodology of this project is found in chapter 3 of the appendix report. To provide background knowledge of the project a literature review of SNOMED CT was conducted. Chapter 4 of the appendix report clarifies the methodology used to conduct this literature review, which includes the strategy of literature search, the criteria of relevance, how literature was evaluated according to methodological quality, and how the included literature was categorised.

2.1 Materials

2.1.1 The dataset

The dataset provided for this project included 941 patients. For each patient 72 DEs were collected. Chapter 5, section 5.1 of the appendix report provides a detailed description of the dataset, while chapter 5, section 6 of the appendix report provides a complete overview of the DEs of the dataset.

The patients included in the dataset had at least a single syncope episode or near syncope episode. For that reason the patients underwent a diagnosing commitment to diagnose the underlying cause of the syncope episodes. The diagnosing commitment included a table tilt test. Most patients were prescribed nitroglycerine to provoke a syncope episode during the table tilt test.

The dataset consisted of demographic data, information about known disorders and medication which may increase the risk of developing syncope episodes, values of pulse- and blood pres-

sures measured during the table tilt test, and the final syncope diagnose. The content of the dataset was information which may be found in the patient's EHR.

To anonymise the data personal information was deleted.

The dataset was stored in a spread sheet ordered with columns of DEs each with rows of data values. The data type of the data values of the DEs were either numerical elements or coded values (multiple categories). Thus, the dataset contained data which was classified to a proprietary, predefined classification. Figure 1 shows a screenshot of a segment of the dataset. Each DE had one data value for each patient. When the number of a specific value found in the dataset for a given DE, the number of samples is specified. If the value was not recorded for a patient the value was "#NUL!". E.g. if the tilt table test was only performed without nitroglycerin no blood pressures or pulse rates were collected for the tilt table test with use of nitroglycerin. But #NUL!-values also appear in cases where no logic explanation for not collecting that information can be found.

If a data value was not recorded the specified value of the dataset was "#NUL!". Missing data values were also found in cases where no obviously, logic explanation appeared. Thus, the reason why data was missing was not specified explicitly and it was not possible to determine if a data value was "missing" or "irrelevant to collect". The lack of explicit contextual knowledge decreased the data completeness and the data quality.

Even though the dataset included a large number of DEs, the level of detail for each DE was low. E.g. most parameters of medication and diagnoses only contained values as "yes"/"no"/"unspecified" which may be considered as the lowest possible level of detail.

2.1.2 Software

For mapping the dataset to SNOMED CT the Healthterm Browser and CliniClue Xplore version 2012.8.0270 were used. The Healthterm Browser was used with the Danish version of SNOMED CT, edition 2013-01-31. CliniClue Xplore was used along with the international version of SNOMED CT, edition 2013-07-31. The Healthterm Browser provided precise translation of some Danish terms which were difficult to translate into English. CliniClue Xplore was used as the primary mapping tool for this project.

Microsoft excel was used to view the syncope dataset, to identify possible data values of DEs, and to count the number of each type of data value.

2.2 Mapping the dataset to SNOMED CT

Mapping the dataset to SNOMED CT was conducted according to a mapping procedure which used a set of quality criteria to refine the mapping. Prior to the mapping of the dataset all DEs of the dataset were divided into groups of similar DEs.

[1]s' mapping guideline was related to the clinical care process where information is recorded according to clinical findings and procedures. E.g. recording of diagnoses, procedural statuses and supplemental information supplemental text fields. The mapping guideline only considered mapping of different data types according to diagnoses and procedures. An Organising Element (OE) cannot always be a procedure or a clinical finding, if the DEs should become subtypes of the OE - this may only be true for EHR-information. In other contexts, as in clinical research, other types of OEs exists.

This project considered mapping of a research dataset which contained 72 DEs which were not ordered into any categories or groups. Further, the dataset included DEs of medication and data values of negation which [1]s' mapping guideline did not support. Therefore, it was necessary to extent [1]s' mapping guideline to be able to use it for mapping of research data.

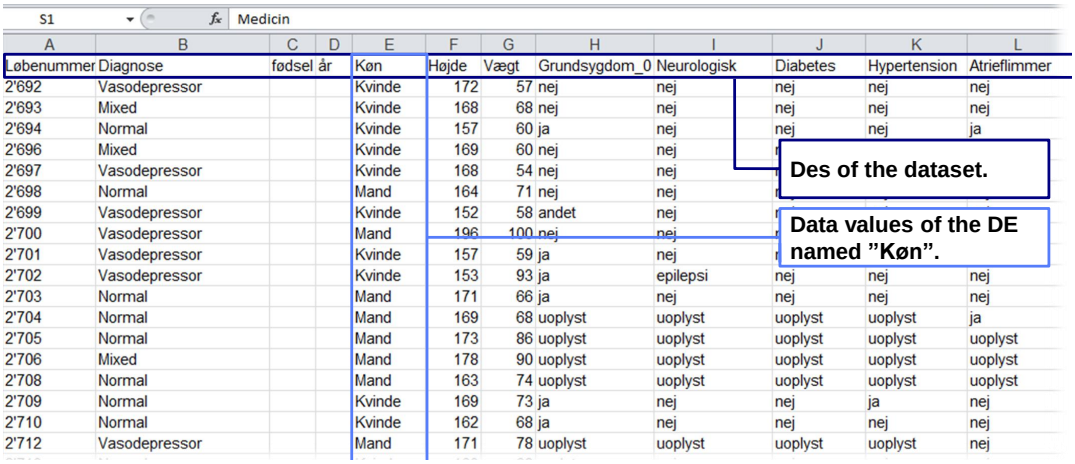
This extension of the mapping guideline included:

- Guidelines of how to group DEs.
- Guidelines of how to map medication to SNOMED CT.
- Guidelines of how to map specific data values as negation, date (month and year), and time.

To develop a methodology which provided guidelines of these three items a purpose and a set of requirements of the mapping were formulated. These are specified in the appendix report, chapter 5, section 5.3 and section 5.4, respectively.

2.2.1 The mapping procedure

The mapping procedure used in this project was based on the mapping procedure developed by [1]. Their mapping procedure was adapted to the objective of this project. The objective of the study



Løbenummer	Diagnose	fødsel år	Køn	Højde	Vægt	Grundsygdom_0	Neurologisk	Diabetes	Hypertension	Atrieflimmer
2'692	Vasodepressor		Kvinde	172	57	nej	nej	nej	nej	nej
2'693	Mixed		Kvinde	168	68	nej	nej	nej	nej	nej
2'694	Normal		Kvinde	157	60	ja	nej	nej	nej	ja
2'696	Mixed		Kvinde	169	60	nej	nej			
2'697	Vasodepressor		Kvinde	168	54	nej	nej			
2'698	Normal		Mand	164	71	nej	nej			
2'699	Vasodepressor		Kvinde	152	58	andet	nej			
2'700	Vasodepressor		Mand	196	100	nej	nej			
2'701	Vasodepressor		Kvinde	157	59	ja	nej			
2'702	Vasodepressor		Kvinde	153	93	ja	epilepsi	nej	nej	nej
2'703	Normal		Mand	171	66	ja	nej	nej	nej	nej
2'704	Normal		Mand	169	68	uoplyst	uoplyst	uoplyst	uoplyst	ja
2'705	Normal		Mand	173	86	uoplyst	uoplyst	uoplyst	uoplyst	uoplyst
2'706	Mixed		Mand	178	90	uoplyst	uoplyst	uoplyst	uoplyst	uoplyst
2'708	Normal		Mand	163	74	uoplyst	uoplyst	uoplyst	uoplyst	uoplyst
2'709	Normal		Kvinde	169	73	ja	nej	nej	ja	nej
2'710	Normal		Kvinde	162	68	ja	nej	nej	nej	nej
2'712	Vasodepressor		Mand	171	78	uoplyst	uoplyst	uoplyst	uoplyst	nej

Figure 1. *A screenshot of the dataset. The dataset included 72 DEs of 941 patients.*

by [1] was to develop a mapping guideline for mapping EHR-template interface terms to SNOMED CT. Thus, the resulting product was a set of mapping guidelines which were developed and refined by iterations of mapping EHR-interface terms to SNOMED CT. In this project the objective was to use a mapping guideline for mapping a research dataset to SNOMED CT. Thus, the resulting product in this project was a research dataset which was mapped to SNOMED CT. The use of an iterative mapping process where refinements are conducted if the mapping does not fulfil a set of quality criteria where inspired from the mapping procedure used by [1]. [1] used a set of quality criteria to develop and refine their mapping guidelines. This project used at set of quality criteria to evaluate the quality of mapping, to ensure consistent mapping. A set of quality criteria were used because the mapping guideline is not well documented yet and it is not yet investigated if the mapping guideline is applicable for mapping research data to SNOMED CT.

Figure 2 shows a flowchart of the mapping procedure used in this project. The mapping procedure was performed as an iterative process divided into 3 steps:

1. Grouping the DEs of the dataset into groups.
2. Each group of DEs were mapped to SNOMED CT. This step was an iterative procedure.
3. Refinements were conducted until the stack of all mapped groups fulfilled a set of quality criteria. This step was also an iterative procedure.

Other studies have focused on making mapping more efficient, by semi-automatic lexical matching and these approaches *are* relevant since mapping clinical information to SNOMED CT is a time-consuming procedure and migration of HISs to SNOMED CT requires mapping of numerous of clinical terms. Since this project focused on *how* research data should be mapped to SNOMED CT and the number of DEs to be mapped were limited it was decided to conduct the mapping manually. Further, it was considered that it would take more time to develop a lexical matching algorithm than conducting mapping manually.

2.2.2 Grouping the data elements

In the research approach of [1] the DEs were naturally divided into groups by the OEs of the EHR-templates. One of the goals of this mapping guideline was to create subset clusters, where the DEs of a group were mapped to subtype descendants of the OE. Therefore, this guideline used these natural groups of DEs and OEs to define that OEs and corresponding DEs should be mapped to the same subtype hierarchy. [1]

The dataset provided for this project did not include a template with OEs which grouped the DEs. Therefore, prior to the mapping of the dataset all DEs of the dataset were divided into groups of similar DEs. Further, [10] have also had good experiences with grouping.

Grouping of the DEs was performed to facilitate consistent mapping. Especially for DEs which required post-coordination grouping was considered to improve consistent use of post-coordination. Further, grouping should make it easier to add new DEs (and more details) to the

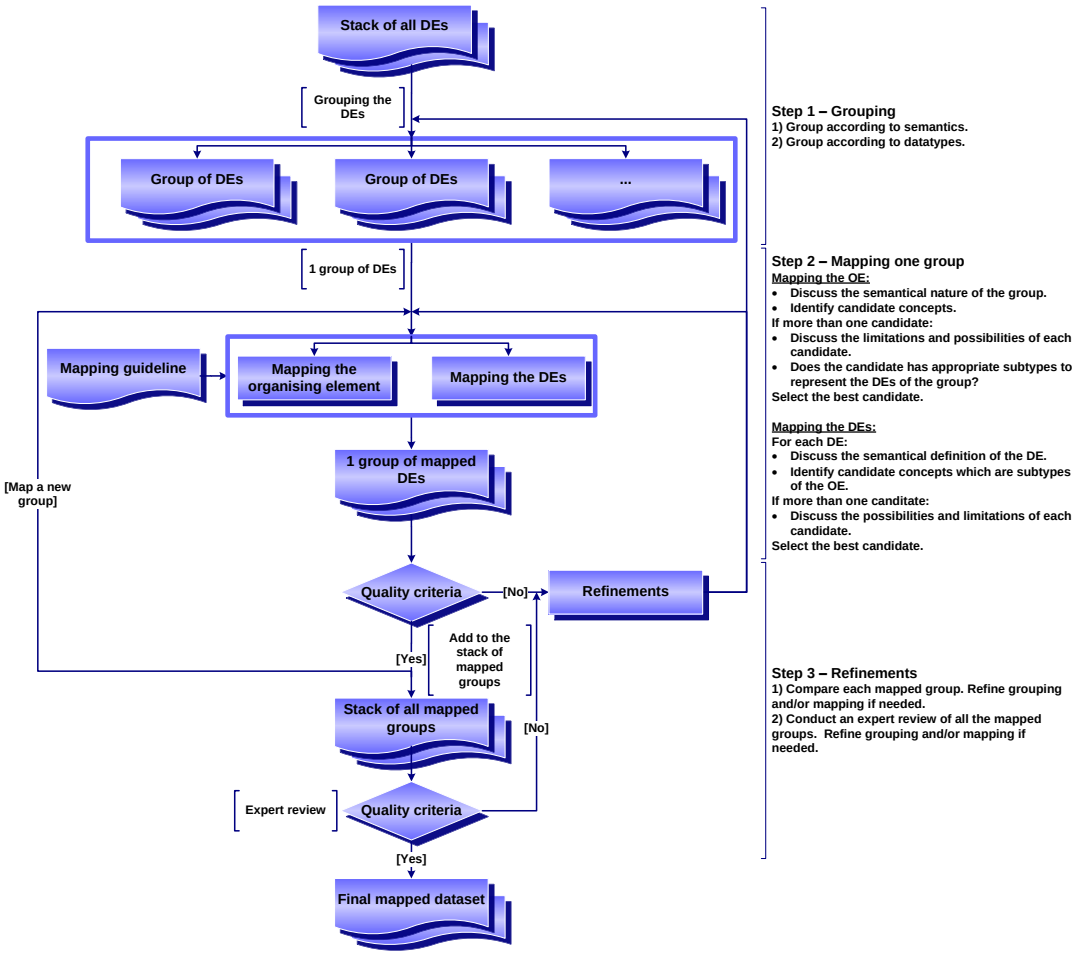


Figure 2. *(This figure is based on an adapted version of the mapping procedure by [1].)*
Mapping the dataset to SNOMED CT was conducted as an iterative process divided into 3 steps; 1) Grouping the DEs, 2) mapping of each group of DEs, and 3) evaluation of all the mapped DEs according a set of quality criteria.

dataset. E.g. if a new diagnosis was to be added an existing group of diagnoses would define how the new diagnosis DE should be mapped to SNOMED CT.

In some cases grouping may be used to apply unambiguous contextual knowledge to a DE, by mapping the DE to the concept of the group name and represent the data values by the group name qualified by a qualifier value. These cases could be DEs which are difficult to map due to lack of appropriate SNOMED CT concepts in SNOMED CT.

OEs which creates post-coordination by combination should be avoided since combined OEs cause a less precise description of the content of appurtenant text fields [1].

The criteria for appropriate grouping were:

1. Groups which consisted of semantically similarly data parameters.
2. Groups which consisted of similar data types.
3. Groups which could be described by a header only consisting of 1 SNOMED CT concept.

Thus, an appropriate group of DEs was e.g. all “disorders”, while a poorly defined group was e.g. “referral diagnosis and final diagnosis”.

2.2.3 Quality criteria

The quality criteria used for this project were formulated based on theory of [2], guidelines of [1], and recommendations of [7].

It was strived to formulate quality criteria which fulfilled that:

- Each quality criterion should be specific and measurable to unambiguously determine if the quality criterion was fulfilled or not.
- The quality criteria should altogether provide an unambiguous selection of SNOMED CT concepts for mapping of DEs to ensure consistent mapping.

The quality criteria formulated for this project were:

QC.1 Fully defined concepts are preferred if possible.
QC.2 Each DE should be represented by a small subset cluster which is as small as possible. Thus, the DE should be supertype parent of the data values.
QC.3 Each group of DEs should be represented by a subset cluster where the OE is the supertype parent of the DEs.
QC.4 Loss of information during mapping to SNOMED CT should be avoided.
QC.5 Pre-coordinated concepts should be selected prior to post-coordinated expressions. If the pre-coordinated concept is primitive post-coordination may be used instead.
QC.6 Post-coordination should be used consistently.

Chapter 8 of the appendix report provides a substantiation of why each quality criterion was formulated.

3. Results

The results are divided into two parts:

- Grouping.
- Mapping - Selection of SNOMED CT concepts.

In the first part of the results the DEs of the dataset were grouped. In the second part of the results candidate concepts for mapping DEs were identified and it is discussed which opportunities and limitations each set of candidate concepts provided. In the third part of the results the final sets of selected SNOMED CT concepts are presented and it is discussed which opportunities and limitations this selection of SNOMED CT provided to the dataset.

3.1 Grouping of the data elements - Results

The DEs of the dataset were divided into the following groups:

- Referral information.
- Diagnoses.
- Miscellaneous.
- Medication.
- Demographic information.
- Table tilt test.

Referral information: Included information about the referring department and the referral diagnosis for which the patient was referred to a diagnosing commitment of syncope.

Diagnoses: Includes information about all the patient's known disorders inclusive specification of the syncope diagnose if the patient was diagnosed during the diagnosing commitment.

Miscellaneous: Included three DEs which did not fit into any other group. This group may has to be redefined or be divided into single DEs to fulfil the grouping criteria specified in the Method, section 2.2.2.

Medication: Includes information about the patient's medication status.

Demographic information: Includes information about the patient's weight, height, age, gender, id-number within the dataset etc.

Table tilt test: This group included all the DEs of the table tilt test e.g. pulse, blood pressure, and supplementary information about the table tilt test.

3.1.1 Delimitation of mapping the dataset

From the dataset the two groups:

- Diagnoses.
- Medication.

were selected and mapped to SNOMED CT. These two groups of DEs were selected since they covered two important important topics of the dataset. Therefore, these two groups of DEs were considered to represent important mapping results according to mapping research data by use of [1]'s mapping guideline.

3.2 Mapping of diagnoses

3.2.1 Data values found in the dataset

The DEs within the group of diagnoses included diagnoses of known disorders. The data values of DEs from the group of diagnoses were:

- Yes. (Or specification of the exact diagnose)
- No.
- Other. (Only for the DE number 4 named

"Grundsygdom_0")
- Unspecified.
- #NUL.

The data type of the DEs of diagnoses may be considered to be list items. The data values of "3) Diagnose" and "5) Neurologisk" were different from the remaining DEs of diagnoses, since the given disorder(s) was/were always specified. For these 2 DEs no data values of "Yes" were found. Further, for "3) Diagnose" no data values of "unknown" or "unspecified" were found. Thus, it was assumed that the syncope diagnose should always be specified, else the contextual reason why syncope was not diagnosed should be specified. E.g. if the diagnosing commitment was not accomplished.

The data values of "3) Diagnose" were:

- The diagnose(s).
- Normal. (Interpreted as the patient was not diagnosed to suffer from syncope)
- Not possible to conduct a conclusion of the diagnose.
- The diagnosing commitment was not accomplished.

The data values of "5) Neurologisk" were:

- The diagnose(s).
- No.
- Unspecified.
- #NUL.

The possible values to select may be configured in a CIS. Thus, all DEs were mapped similarly, to facilitate similarly selection of data values for all DEs within the diagnoses group.

3.2.2 Requirements of further functionality for future purposes

For diagnose it was considered to be of interest to facilitate specification of:

- The given disorder which the patient suffers from.
- More than one disorder for each DE.
- Other types of disorders which were not found in the dataset.
- The data values with different levels of granularity, dependent on relevance and the information available in the EHR-system. (E.g. a group of disorders or the given exact diagnose)

3.2.3 Mapping the OE of diagnoses

Lack of subtype relationships between the OE and the result fields of that has been discussed to be problematic. [1] For that reason candidates for the OE were identified with both the OE and the DEs of the OE in mind.

Depending on the context a diagnose may be the referral diagnose, action diagnose, the final diagnose etc. This may change in different contextual settings. An action diagnose could become the final diagnose, and a clinical finding in a department could become the referral diagnose in another department. For that reason it did not make sense to specify the contextual type of diagnose.

A result or a finding from a clinical observation is considered to be a clinical finding [7]. According to [7] a clinical finding should be mapped to a concept from the |Clinical finding (Top-level concept)| subtype hierarchy. This recommendation agree with [1]'s mapping guideline. The disorders specified in the dataset provided for this project may be a result of a data extraction of diagnoses (SKS-codes) found in the EHR. Therefore, it was considered to be appropriate to select candidate concepts from the |Clinical finding (Top-level concept)| subtype hierarchy.

Candidates for mapping the OE of diagnoses

Table 2 shows the identified candidate concept for mapping the OE of diagnoses to SNOMED CT. Since it was obvious which SNOMED CT concept to select only a single candidate was identified.

The advantages, disadvantages, and consequences of selection of this candidate are clarified in the following. Advantages are marked with "**+**" while disadvantages are marked with "**-**".

|Disorder of body system (Clinical finding)|

+ |Disorder of body system (Clinical finding)| provided a subtype relationship between the OE and the DEs of the group.
+ |Disorder of body system (Clinical finding)| supported specification of diagnoses with optional level of granularity, since this concept both includes subtype descendants which specifies groups of disorders and exact diagnoses.
+ |Disorder of body system (Clinical finding)| facilitated adding of more DEs to specify any diagnose.
+ The concept was fully defined.

OE	SNOMED CT concept	Concept ID	Defined
Diagnoses	|Disorder of body system (Clinical finding)|	362965005	Fully specified

Table 1. *Candidates for mapping of the OE named diagnoses.*

- The concept was coarse grained, thus creating a large subset cluster.

Selection of SNOMED CT concepts for the OE of diagnoses

|Disorder of body system (Clinical finding)| provided selection of subtype descendants for mapping of the DEs of diagnoses. Therefore consistent mapping was provided and it was possible to create subset clusters of DEs which were subtype descendants of the OE. |Disorder of body system (Clinical finding)| provided sufficient subtype descendants to map all DEs and data values of diagnoses without loss of information. Further, the selected SNOMED CT concept was fully defined.

3.2.4 Mapping the DEs of diagnoses

Selection of SNOMED CT concepts for the DEs of diagnoses is discussed below. DEs which were obvious to map are not clarified.

Diagnose (3)

The data values of "3) Diagnose" found in the dataset were different types of syncope and the situational context.

Grundsygdom_0 (4)

This DE was interpreted as a residual class which purpose was to specify any other disorders. For that reason the DE was mapped as a residual class according to the mapping guideline of [1].

Neurologisk (5)

The data values of "5) Neurologisk" included some specific disorders which did not share a common pathophysiological nature, thus were not semantically likely. The disorders were mental disorders, degenerative disorders, brain disorders, disorders of CNS and/or the nervous system not specifically related to CNS. To ensure a corresponding SNOMED CT concept for each data value the least common parent |Disorder of nervous system (Clinical finding)| was selected. This SNOMED CT concept was coarse-grained and had several other subtype descendants irrelevant of the application of this project.

Table 2 provides an overview of the selected SNOMED CT concepts for mapping of the DEs. Table 3 provides an overview of how the data values of the dataset should be mapped to SNOMED CT.

3.2.5 Overview of the subtype relationships between the mapped DEs of diagnoses

Figure 3 shows an overview of the subtype relationships between the SNOMED CT concepts selected for the mapping of the DEs of diagnoses. The OE of diagnoses is mapped to |Disorder by body system (Clinical finding)|. Remark that the DEs of diagnoses are inherited from the three SNOMED CT concepts |Disorder of cardiovascular system (Clinical finding)|, |Disorder of nervous system (Clinical finding)|, and |Disorder of endocrine system (Clinical finding)|. These SNOMED CT concepts forms three main groups. The DEs of diagnoses does not have the same level of granularity and some DEs are subtype descendants of others. E.g. |Cerebrovascular accident (Clinical finding)| is a subtype descendant of |Disorder of nervous system (Clinical finding)|. Some DEs have multiple parents. E.g. |Cerebrovascular accident (Clinical finding)| is both inherited from |Disorder of nervous system (Clinical finding)| and |Disorder of cardiovascular system (Clinical finding)|.

The |Clinical finding (Top-level concept)| subtype hierarchy of SNOMED CT was generally well defined and included several subtype descendants of which many were fully defined. Therefore, the |Clinical finding (Top-level concept)| subtype hierarchy provided high expressiveness. Further, the |Clinical finding (Top-level concept)| subtype hierarchy of SNOMED CT provided different dichotomisations, e.g. dichotomisation according to body site, body system, and according to type of disease.

3.3 Mapping of medication

3.3.1 Data values found in the dataset

The medication group of the dataset included drug therapy of cardiovascular disorders, neurological disorders, psychiatric disorders, and di-

No.	DE	SNOMED CT concept	Concept ID	Defined
3	Diagnose	|Syncope (Disorder)|	Primitive	
4	Grundsygdom_0	|Disorder by body system (Clinical finding)| + |Other (Qualifier value)|	362965005 + 74964007	Fully defined + Primitive
5	Neurologisk	|Disorder of nervous system (Clinical finding)|	118940003	Fully defined
6	Diabetes	|Diabetes mellitus (Clinical finding)|	73211009	Primitive
7	Hypertension	|Hypertensive disorder (Clinical finding)|	38341003	Primitive
8	Atrieflimmer	|Atrial fibrillation (Clinical finding)|	49436004	Fully defined
9	Iskaemisk_hjertesygdom	|Ischemic heart disease (Clinical finding)|	414545008	Fully defined
10	Cerebralt_insult	|Cerebrovascular accident (Clinical finding)|	230690007	Primitive
11	Lavt_blodtryk	|Low blood pressure (Clinical finding)|	45007003	Primitive

Table 2. *Selected SNOMED CT concepts for mapping of the DEs for diagnoses.*

Data Value	SNOMED CT concept	Concept ID	Defined
• Yes	(<<) |Disorder of body system (Clinical finding)|	362965005	Fully defined
• No	(==) |Disorder of body system (Clinical finding)|	362965005	Fully defined
• Unspecified	(==) |Disorder of body system (Clinical finding)|	362965005	Fully defined
• Other	(<<) |Disorder of body system (Clinical finding)|	362965005	Fully defined

Table 3. *Specification of the allowed SNOMED CT concepts for each data value of DEs within the group of diagnoses.*

abetes. Further, general medication and diuretic drug therapy were included. The data values which were generally presented within this group of the dataset were:

- Yes.
- No.
- Unspecified. (Not all types of medication)
- #NULL!. (Not all types of medication)

The data type of the DEs of medication may be considered to be list items.

3.3.2 Requirements of further functionality for future purposes

For medication it was considered to be of interest to facilitate specification of:

- More types of medical therapy, which were not included in the dataset provided for this project.
- All prescribed pharmaceutical products and from that count the number of prescribed pharmaceutical products. Further, it may provide more details to the research study.
- Dose of prescribed pharmaceutical products.
- More than a single pharmaceutical product for one DE.

Interoperability with current systems and standards within medication

To enhance applicability and interoperability with current systems and standards these were considered. Currently the The Anatomical Therapeutic Chemical (ATC) classification system is used in DK and the Shared Medication Record (FMK, det Faelles Medicin Kort) is about to be implemented on a national basis [15, 16].

The ATC classification system is an international classification system of pharmaceutical substances. Maintenance is performed by WHO.

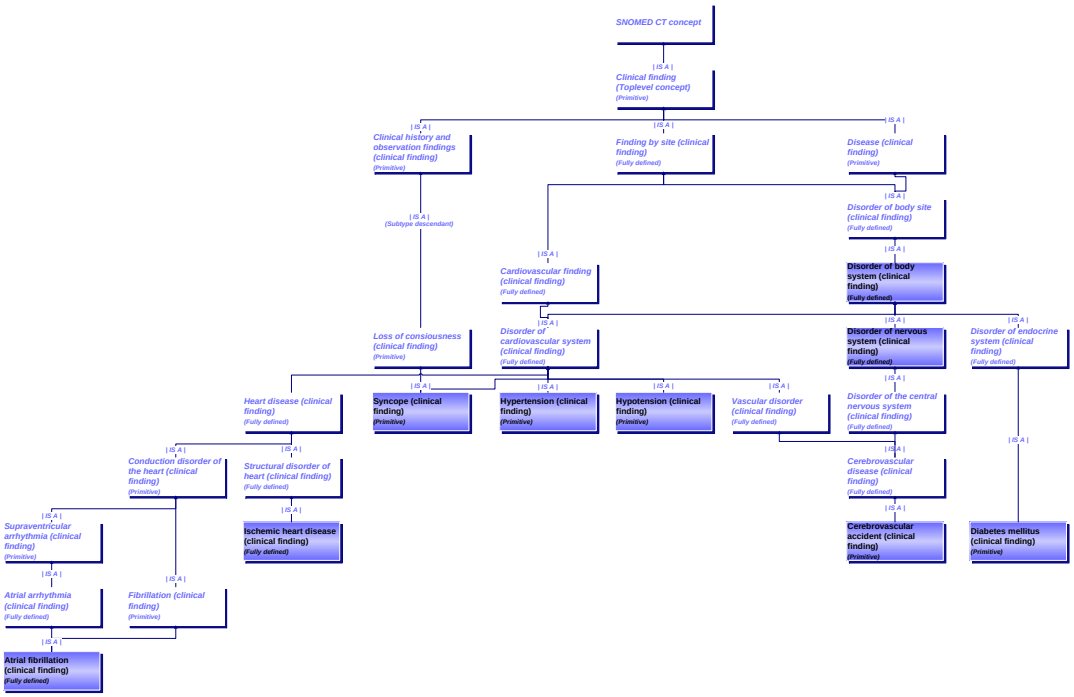


Figure 3. *Overview of the subtype relationships between the DEs of the diagnoses group. The figure shows a subset of relevant SNOMED CT concepts. SNOMED CT concepts which represent a DE are depicted with blue background colour. The SNOMED CT concepts which are included in this figure have several subtype descendents which are not depicted. Some supertype ancestors are also left out. (A3 version of the figure is found at the back of this article in section 5.2).*

Classification of medicine in SKS is based on a Danish translation of the ATC-classification system. The ATC-classification system is developed to provide drug utilisation research. The ATC-classification system only specifies the chemical constituent of a pharmaceutical product along with a defined daily dose (DDD). The DDD specifies the daily assumed average maintenance dose of a chemical constituent. [15]

FMK is a national service which provides sharing of a patient's current medical therapy. When FMK is fully implemented all CIS in general practice and in hospitals will retrieve and report information about all patients' medication in DK to this national service. In that way access to up-to-date information about medication is provided nationally [16]

Since FMK is capable of handling specification of therapeutic indication, dosage, and pharmaceutical product [16], it may not be necessary to let this functionality be handled by SNOMED CT in order to facilitate specification of medication with a higher level of granularity than specified in the current dataset and still facilitate interoperability with national standards. But only SNOMED CT facilitates interoperability on an international level. Linkage between FMK and SNOMED CT may be possible by linkage between chemical constituents in SNOMED CT and corresponding ATC-codes. Specification of medication in FMK also includes ATC-codes [16].

3.3.3 Mapping the OE of medication

Candidates for the OE were identified with both the OE and the DEs of the OE in mind.

According to [1] OEs should be mapped to a SNOMED CT concept from either the |Clinical finding (Top-level)| subtype hierarchy or the |Procedure (Top-level)| subtype hierarchy. But they did not discuss mapping of medication in their guideline. A physician may observe/ascertain that a patient takes a given type of medication or the physician may prescribe medication to a patient. For that reason "medication" may be considered as either an observation or a procedure depending on the contextual situation in which the information is recorded, with the prescribed types of medication/pharmaceutical products as

the central subject of interest. In the context of this project no medication was prescribed. The information of medication may be obtained by a look-up in the patients' medication record. Thus, medication might be considered as an observation or a clinical finding. It is recommended to use the |Procedure (Top-level concept)| subtype hierarchy to specify the act of administrating medicine to a patient [17, sec. 6.2.1]. The intention in this project was not to facilitate documentation of the act of prescribing medicine to a patient, but to retrieve information about a patient's current medical therapy.

The |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchy provides specification of pharmaceutical products [7]. For that reason selection of a SNOMED CT concept from the |Clinical finding (Top-level concept)|- or the |Procedure (Top-level concept)| subtype hierarchy for mapping the OE of medication did not create subtype relationships to groups of pharmacological products or the given pharmacological products used for the medical therapy.

Due to these conflicting recommendations of [1] and [17] it was investigated if the OE of medication could be mapped by selection of a SNOMED CT concept from the |Substance (Top-level concept)|- or the |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchy. Searching for candidates was conducted within the top-level hierarchies: |Procedure (Top-level concept)|, |Pharmaceutical/biologic product (Top-level concept)|, and |Substance (Top-level concept)|.

Candidates for mapping the OE of medication

Table 4 shows the identified candidates of the SNOMED CT concepts for mapping the OE of medication. The advantages, disadvantages, and consequences of selection of each candidate are clarified in the following. Advantages are marked with "**+**" while disadvantages are marked with "**-**".

|Drug or medicament (Substance)|

+ |Drug or medicament (Substance)| included subtype descendants to specify any active ingredient of a drug or any group of active ingredients. Therefore |Drug or medicament (Substance)| supported optional level of granularity.

+ |Drug or medicament (Substance)| provided a subtype relationship between the OE and the DEs of the group.

- |Drug or medicament (Substance)| did not facilitate specification of dose and dose form within SNOMED CT by use of subtype descendants, because this SNOMED CT concept does not has subtype descendents or defining relationships which specifies dosage.

- |Drug or medicament (Substance)| is a subtype child to the top-level concept Substance. Thus, the concept was coarse-grained and selection of this concept would create a large subset cluster.

- |Drug or medicament (Substance)| did not provides specification of a combinatorial pharmaceutical product by use of a single SNOMED CT concept.

- Specification of explicit context within SNOMED CT was not facilitated by |Drug or medicament (Substance)|.

- The |Substance (Top-level concept)| subtype hierarchy is a supportive subtype hierarchy. Thus, it is intended to support a main hierarchy, and can only become the range of a post-coordination.

- The concept was primitive.

|Pharmaceutical/biologic product (Top-level concept)|

+ |Pharmaceutical/biologic product (Top-level concept)| provided a subtype relationship between the OE and the DEs of the group.

+ |Pharmaceutical/biologic product (Top-level concept)| supported specification of medication with optional level of granularity.

+ |Pharmaceutical/biologic product (Top-level concept)| facilitated specification of dose and dose form of medical products within SNOMED CT.

+ |Pharmaceutical/biologic product (Top-level concept)| includes pre-coordinated SNOMED CT concepts which facilitates specification of combinatorial pharmaceutical products by a single SNOMED CT concept.

- |Pharmaceutical/biologic product (Top-level concept)| is a top-level concept. Thus, the concept is very coarse-grained.

- Specification of explicit context within SNOMED CT was not facilitated by |Pharmaceutical/biologic product (Top-level concept)|.

- The concept was primitive.

Figure 4 shows an example of what possibilities and limitations the |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchy

OE	SNOMED CT concept	Concept ID	Defined
Medication	|Drug or medicament (Substance)|	410942007	Primitive
	|Pharmaceutical/biologic product (Top-level concept)|	373873005	Primitive
	|Medical therapy (Procedure)|	243121000	Primitive
	|Drug therapy (Procedure)|	416608005	Fully specified

Table 4. *Candidates for mapping the OE of medication.*

provides and how it is related to the |Substance (Top-level concept)| subtype hierarchy.

In SNOMED CT pharmaceutical products and their chemical constituents are divided into two subtype hierarchies. The two subtype hierarchies provide an explicit distinction between pharmaceutical products and the chemical constituents of the pharmaceutical products. The relationship between drug products and the chemical constituents is defined by the defining relationship |Pharmaceutical/biologic product (Top-level concept)| |Has active ingredient (Linkage concept)| |Substance (Top-level concept)|. [7]

Use of specific brands of pharmaceutical products may differ among countries. Therefore, a part of the |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchy of SNOMED CT is country specific. The international version of SNOMED CT includes concepts of Virtual Therapeutic Moieties (VTM) which specifies a therapeutic drug without any dose or dose form. |Nitrogen (Product)| and |Acetaminophen (Product)| are examples of VTMs. VTMs have similar subtype descendants which are pre-coordinated with dose and dose form. These subtype descendants are denoted Virtual Medicinal Products (VMP). [7]

The country specific specific extension includes corresponding Actual Medicinal Products (AMP). Each VMP in the international version of SNOMED CT is mapped to a corresponding AMP in the country specific extension of SNOMED CT. [7]

The purpose of the international version of SNOMED CT is to act as a standard or reference model, thus translation of pharmacological content may be conducted without loss of the essential meaning, namely the chemical constituent(s), the dose, and the dose form. [7]

In figure 4 the VMP |Paracetamol 500 mg tablet (Product)| in the international version of SNOMED CT is mapped to the corresponding AMP, |Panodil 500 mg tablet (Product)|, in the Danish extension. Other AMPs as Pinex or Pamol could also be represented in the Danish extension if the |Product (Top-level concept)| subtype hierarchy had a higher level of maturity.

Dose form is provided explicitly by the defining attribute |Has dose form (Linkage concept)|. [17] Therefore, it is possible to determine the dose form of the VMPs (which are pre-coordinated) from their definition by computer processing. Unfortunately, SNOMED CT does not provides expression of numerical numbers and no defining relationships defines the dose of a VMP. Therefore, it is not possible to determine the dose of the VPMs from the defining relationships of the VPM by computer processing.

|Drug therapy (Procedure)|

+ |Drug therapy (Procedure)| has subtype descendents of which more of them are fully specified. Thus, it was possible to represent different types of medical treatment by use of subtype descendants of |Drug therapy (Procedure)|.
+ |Drug therapy (Procedure)| provided refinements by use of the attribute |Direct Substance (Linkage concept)| to specify the pharmaceutical product. Although the allowed range of |Procedure (Top-level concept)| |Direct Substance (Linkage concept)|, defined by the SNOMED CT Concept Model, both includes subtype descendants of |Substance (Top-level concept)| and |Pharmaceutical/biologic product (Top-level concept)|, the subtype descendants of |Drug therapy (Procedure)| *only* have defining relationships with subtype descendants of |Substance (Top-level concept)|.
+ Use of the Situation with explicit context hierarchy provided specification of the data values within SNOMED CT. Thus, the context was applied explicitly and the mapped dataset would get a high degree of independence of a proprietary information model of a CIS.
+ The concept was fully specified.

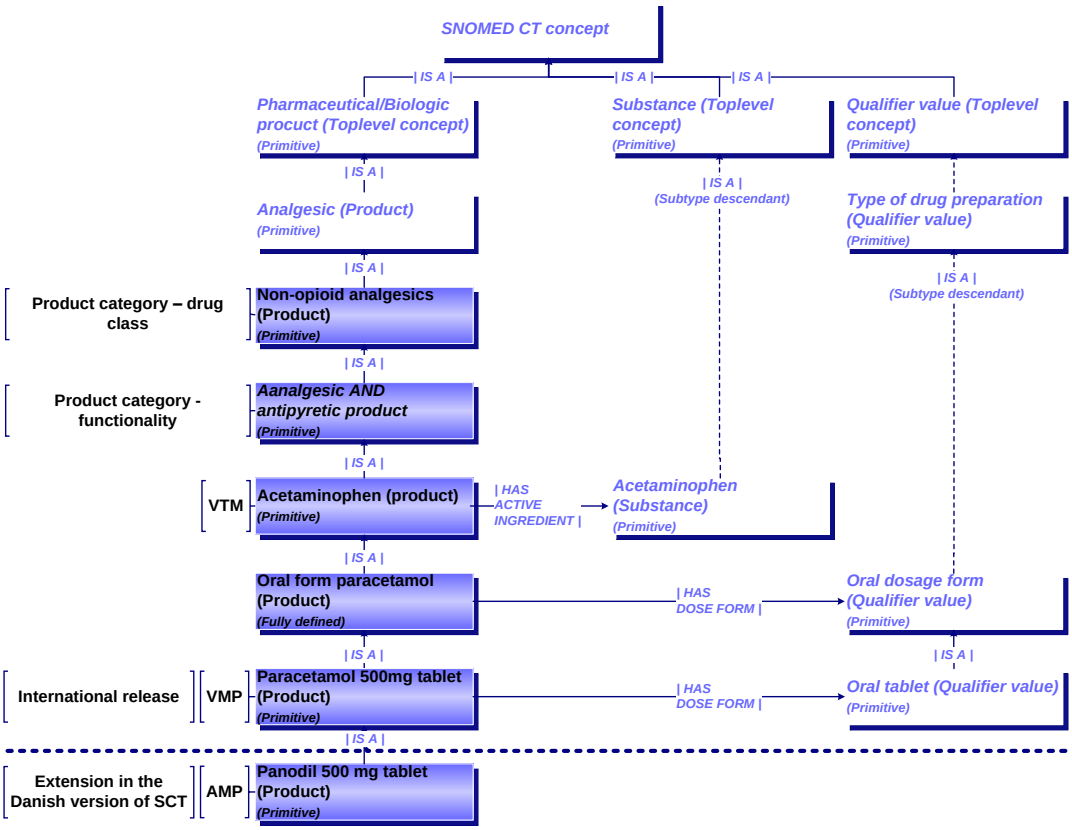


Figure 4. *(Modification of [7]. The example is an adaption of a figure from [7]. The original figure showed an example of the VMP \Alteplase 10mg powder and solvent for injection solution vial (Product)\. Further, the defining relationship of dose form is applied in this figure.)*

Illustration of how pharmaceutical products are modelled in SNOMED CT. This figure shows an example of how the VMP \Paracetamol 500 mg tablet (Product)\ is represented in SNOMED CT.

- |Drug therapy (Procedure)| did not provide specification of the pharmaceutical products (along with dose form) used for the medical treatment, by use of subtype descendants of |Drug therapy (Procedure)|.
- Despite of several relevant subtype descendants of |Drug therapy (Procedure)| these were not sufficient to represent all DEs of medication without significant loss of information. The reason for that was inappropriate dichotomisation of the subtype descendants to map DEs of cardiovascular medication.
- DEs would be represented by post-coordinated expressions.
- Since the context values were qualifier values it may not be possible to search for these during queries, unless the database is designed to support this feature. Therefore it may not be possible to select only those patients who *take* medication or those who do *not take* medication by a query.

Figure 5 shows a list of relevant subtype descendants of |Drug therapy (Procedure)|. From the list it is ascertained that mapping the DEs to this hierarchy, caused a need of subdivision of the DE “18) Hjertemedicin”. Further, the |Drug therapy (Procedure)| did not has any subtype descendants to specify hypotensive therapy. Thus, the DEs “15) Nitroglycerin” and “17) Antihypertensiva” could not be mapped.

Figure 6 shows an example of the possibilities and limitations provided if |Drug therapy (Procedure)| was selected. |Drug therapy (Procedure)| included subtype descendents of more fine-grained types of medical therapy. These SNOMED CT concepts have defining relationships to subtype descendants of the |Substance (Top-level concept)| subtype hierarchy, thus forming subset clusters of substances provided for a given type of medication. The use of SNOMED CT concepts from the |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchy as values for

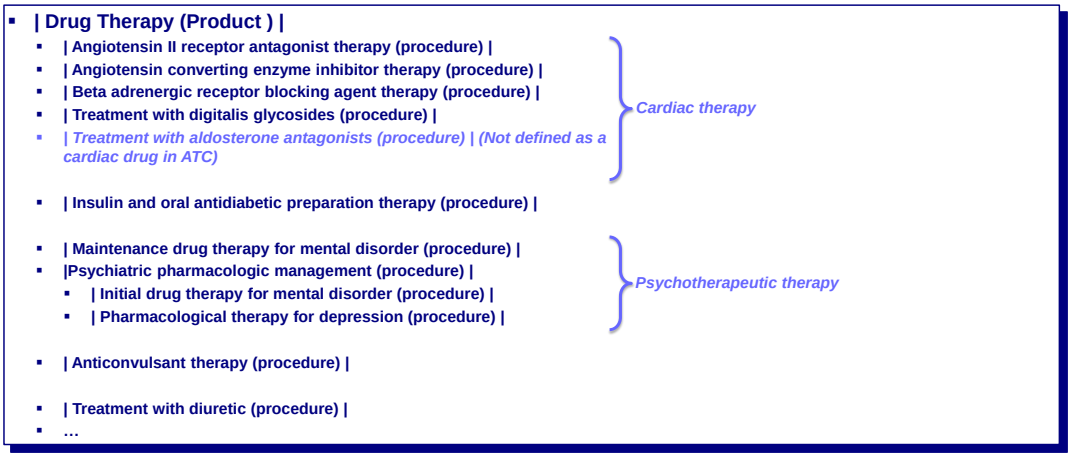


Figure 5. *Listing of relevant subtype descendants of |Drug therapy (Procedure)|.*

the |Direct Substance (Linkage concept)| attribute also allowed by the Concept Model of SNOMED CT, but not recommended by IHTSDO [7].

The defining relationship |Pharmaceutical / biologic product (Top-level concept)| |Has active ingredient (Linkage concept)| |Substance (Top-level concept)| provided dose form and specification of a country specific pharmaceutical products.

The |Direct Substance (Linkage concept)| attribute is provided for cases where the substance / pharmaceutical product used is the object of the procedure. Legal attribute values for the |Direct Substance (Linkage concept)| are concepts from the Substance- and the Pharmaceutical/biologic product subtype hierarchies. Only use of SNOMED CT concepts from the |Substance (Top-level concept)| subtype hierarchy are recommended along with the international version of SNOMED CT. [17]

The Situation with explicit context hierarchy provides explicit specification of the context of the data, which corresponds to the data values of medication, by qualification with a subtype descendant of |Context values for actions (Qualifier value)|.

Figure 7 shows the subtype descendents of |Context values for actions (Qualifier value)| available if the data values of the dataset were mapped to SNOMED CT, by use of the context model of SNOMED CT. From the list it is observed that SNOMED CT includes at least two SNOMED CT concepts for the data value “Yes”, depending on whether the patient had received medication earlier or is still receiving medication. These context values are defined to specify the procedural status. Therefore, it was problematic to use these values to specify if a patient takes medication or not. Thus, the context values of SNOMED CT were not representative for the data values of the current dataset.

To ensure statistical reliability of a research project, it may not be appropriate to design a study protocol with 53 context values (as available in SNOMED CT), since it will require a large number of samples to obtain normal distribution in the dataset. Normal distribution is necessary to say that the samples in the dataset are representative for the general population.

The code in figure 8 shows how selection of |Drug therapy (Procedure)| creates a post-coordinated expression with the data value specified by a qualifier value.

|Medical therapy (Procedure)|

This concept was similar to |Drug therapy (Product)|.

+ |Medical therapy (Procedure)| provided refinements by use of the attribute |Direct Substance (Linkage concept)|. This attribute provided specification of the prescribed medical substances and the pharmacological products along with dose form.
- |Medical therapy (Procedure)| did not include the subtype descendants needed to represent all types of medical therapy found the the dataset provided for this project. Thus, it was not possible to represent the DEs of medication by use of subtype descendants of the |Medical therapy (Procedure)| concept.
- |Medical therapy (Procedure)| did not provide

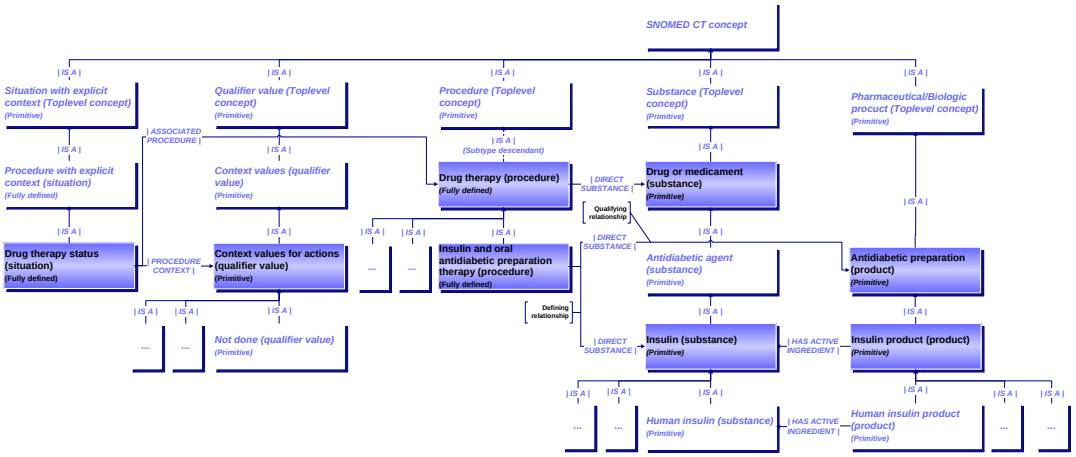


Figure 6. *Example of the limitations and possibilities provided if the OE of medication was mapped to |Drug therapy (Product)| and the DEs were mapped to subtype descendants of |Drug therapy (Product)|. Specification of explicit context within SNOMED CT is provided by qualification with a subtype descendant of |Context values for actions (Qualifier value)|. Note that specification of specific pharmaceutical products is possible by* ***refinement*** *of defining relationships to concepts from the |Substance (Top-level concept)| subtype hierarchy or by* ***qualification*** *with a concept from the |Pharmaceutical/biologic product (Top-level concept)|. (A3 version of the figure is found at the back of this article in section 5.2).*

specification of the pharmaceutical products (along with dose form) used for the medical treatment, by use of subtype descendants of |Medical therapy (Procedure)|.

- The concept was primitive.

Selection of the SNOMED CT concept for mapping the OE of medication

According to quality criteria no. QC.1. only |Drug therapy (Procedure)| was fully defined.

|Medical therapy (procedure)| and |Drug therapy (Procedure)| created smaller delimited subset clusters according to quality criteria no. QC.2. |Pharmaceutical/biologic product (Top-level concept)| and |Drug or medicament (Substance)| created larger and less delimited subset clusters.

QC.3: |Medical therapy (procedure)| and |Drug therapy (Procedure)| provided specification of the DEs (not all of them), by use of subtype descendants. More detailed specification of medical therapy required refinement/qualification, with concepts from the |Substance (Top-level concept)| or |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchy. Only |Pharmaceutical/biologic product (Top-level concept)| and |Drug or medicament (Substance)| included subtype descendants needed to specify medical therapy with optional level of granularity.

Neither |Medical therapy (procedure)| nor |Drug therapy (Procedure)| included sufficient subtype descendents (and dichotomisation of these) to represent all the DEs of medication, without loss of information. Thus, these concepts did not fulfil quality criteria no. QC.4. |Pharmaceutical/biologic product (Top-level concept)| and |Drug or medicament (Substance)| did not cause loss of information.

The main difference between the advantages of |Pharmaceutical/biologic product (Top-level concept)| and |Drug or medicament (Substance)| was that the |Pharmaceutical/biologic product (Top-level concept)| provided specification of dose and dose form of medication. Further, |Pharmaceutical/biologic product (Top-level concept)| included fine-grained pre-coordinated SNOMED CT concepts to represent combinatorial pharmaceutical products by use of a single SNOMED CT concept. |Drug or medicament (Substance)| did not included pre-coordinated fine-grained SNOMED CT concept. Thus, combinatorial pharmaceutical should be specified by one SNOMED CT concept for each active ingredient. The subtype descendants of |Drug or medicament (Substance)| also included non-pharmaceutical products, e.g. non-medical psychedelic drugs. The dichotomisation within the two hierarchies only had minor differences.

Figure 7. *The context values available, if the context model of SNOMED CT was used to specify the data values. The concepts marked with dark blue colour are those which corresponds to the data values found in the dataset. ("Yes", "no", "unspecified") From the list it is observed that SNOMED CT provided several values, of which only few concepts were relevant in a research dataset.*

It was considered to be problematic to use the SNOMED CT concepts from the |Context values for actions (Qualifier value)| to represent the data values of medication. Therefore, the data values were not mapped to SNOMED CT.

|Pharmaceutical/biologic product (Top-level concept)| was chosen since it created a subtype relationships between the OE and the DEs of the group. Further, this concept included all types of medication to provide specification of more types of medication for future purposes.

3.3.4 Mapping of the DEs of medication

Selection of SNOMED CT concepts for the DEs of medication is discussed below. DEs which were obvious to map are not clarified.

Medicin (14)

From the dataset it was observed that the data values of this DE could take the value "Yes", despite all other data values of medication were "No". Thus, it is not clear what kind of medication this data value covers. For that reason it was not possible to map this DE.

Andet_medicin (22)

This DE is a residual category to ensure exhaustiveness. It may serve as a supplementary DE. To facilitate free text this DE was mapped to the OE + a qualifier value according to [1]s' mapping guideline. Further, by selection of a subtype descendant of the OE, it is possible to specify what other other type of medication with optional level of granularity.

Antiepileptika (19)

No SNOMED CT concepts to specify antiepileptic medical therapy, antiepileptic medical substances, or - products were found. In stead a concept which included anticonvulsant drugs was selected, since antiepileptic drugs are a kind of anticonvulsant drugs. Thus the overall semantic meaning was still maintained.

Some anticonvulsant drugs are used for other purposes than epileptic convulsions. E.g. to depress cough or treat arrhythmia. Therefore, the subset cluster of anticonvulsant drugs also included some irrelevant subtype descendants.

Table 5 shows the SNOMED CT concept selected for mapping of the DEs. As shown in table 6 the answer ("Yes/No/Unspecified") should be

```
129125009 | procedure with explicit context | :
   { 363589002 | associated procedure | = ( 1821000052107 | insulin and oral antidiabetic
preparation therapy | :
   363701004 | direct substance | = 67866001 | insulin | )
   , 408730004 | procedure context | = 385660001 | not done |
   , 408731000 | temporal context | = 410512000 | current or specified |
   , 408732007 | subject relationship context | = 410604004 | subject of record |
}
```

Figure 8. *The post-coordinated expression of |Drug therapy (Procedure)|, refined with specification of contextual knowledge and use of medical substance.*

No.	DE	SNOMED CT concept	Concept ID	Defined
14	Medicin	Unable to map		
15	Nitroglycerin	|Nitroglycerin (Product)|	71759000	Primitive
16	Vanddrivende	|Diuretic (Product)|	30492008	Primitive
17	Antihypertensiva	|Hypotensive agent (Product)|	1182007	Primitive
18	Hjertemedicin	|Cardiac drug (Product)|	30067000	Primitive
19	Antidiabetika	|Antidiabetic preparation (Product)|	384953001	Primitive
20	Antiepileptika	|Anticonvulsant (Product)|	63094006	Primitive
21	Psykopharmaka	|Psychotherapeutic agent (Product)|	46063005	Primitive
22	Andet_medicin	|Pharmaceutical/biologic product (Top-level concept)| + |Other (Qualifier value)|	373873005 + 74964007	Primitive + Primitive

Table 5. *Selected SNOMED CT concepts for mapping the DEs of medication. Since it was not possible to interpret the semantic meaning of "'14) Medicin"', this DE was not mapped.*

stored separate from SNOMED CT since data values do not represent general terminological content. If the answer was "Yes" the mapped dataset provides specification of medication with optional level of granularity from the most coarse-grained possible option "Yes/No/Unspecified" to the fine-grained level which includes specification of a given pharmaceutical product along with dose. One or more subtype descendants may be selected to specify one or more medical products.

3.3.5 Overview of the subtype relationships between the mapped DEs of medication

Figure 9 shows an overview of the subtype relationships between the SNOMED CT concepts selected for mapping of the DEs of the medication group. The OE of medication was mapped to |Pharmaceutical/biologic product (Top-level concept)| and the DEs were mapped to more fine-grained SNOMED CT concepts. Remark that the DEs of medication are inherited from the four SNOMED CT concepts |Renal drug (Product)|, |Cardiovascular drug (Product)|, |Antidiabetic preparation (Product)|, and |CNS drug (Product)|. These SNOMED CT concepts forms four main groups. From the figure it is observed that the DEs does not have the same level of granularity and the DE |Nitroglycerin (Product)| is a subtype descendant of |Hypotensive agent (Product)|. Further, |Diuretic (Product)| is both a |Renal drug (Product)| and a |Cardiovascular drug (Product)|. All the selected SNOMED CT concepts are primitive.

No context values are provided to post-coordinate SNOMED CT concepts from the |Pharmaceutical/Biologic product (Top-level concept)| subtype hierarchy, as for SNOMED CT concepts from the |Clinical finding (Top-level concept)|- and |Procedure (Top-level concept)| subtype hierarchy. Therefore, it was not possible to represent the data values of medication by post-coordination of the DEs, since no linkage concepts.

Data Value	SNOMED CT concept	Concept ID	Defined
• Yes	(<<) |Pharmaceutical/biologic product (Top-level concept)|	373873005	Primitive
• No	(==) |Pharmaceutical/biologic product (Top-level concept)|	373873005	Primitive
• Unspecified	(==) |Pharmaceutical/biologic product (Top-level concept)|	373873005	Primitive

Table 6. *Specification of the allowed SNOMED CT concepts for each data value of the DEs within the group of medication.*

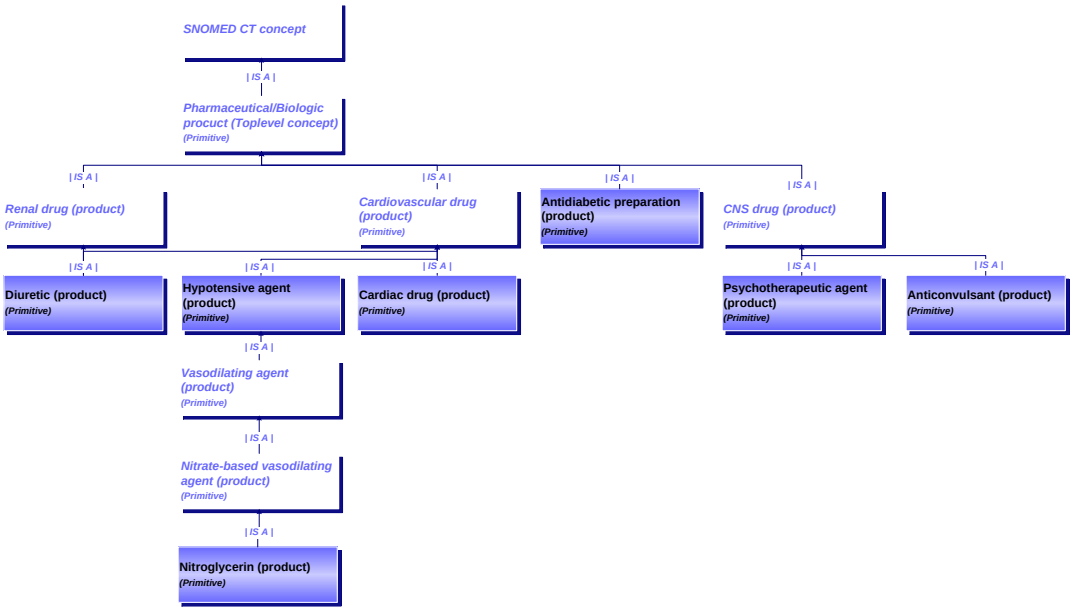


Figure 9. *Overview of the subtype relationships between the DEs of medication. The figure shows a subset of relevant SNOMED CT concepts. SNOMED CT concepts which represent a DE are depicted with blue background colour. The SNOMED CT concepts which are included in this figure have several subtype descendents which are not depicted.*

4. Discussion

4.1 Discussion - What may SNOMED CT add to research data?

4.1.1 Findings which were general for all groups of DEs

Inconsistency and incompleteness of SNOMED CT

As other studies have found incompleteness and inconsistency of SNOMED CT [10, 9, 1, 8, 5], these problems are also experienced in this project. Inconsistency problems include inconsistent naming of SNOMED CT concepts and inconsistency in the hierarchies. E.g. inconsistency of the dichotomisation of the subtype descendants of |Drug therapy (Procedure)| was experienced. Some of the subtype children of |Drug therapy (Procedure)| were fine grained (e.g. the SNOMED CT concepts which specified a type of cardiac medication) while others were coarse grained (E.g. |Anticonvulsant therapy (Procedure)|).

Incompleteness of SNOMED CT includes lack of- SNOMED CT concepts, defining- and qualifying relationships, support of negation, and fully defined intermediate SNOMED CT concepts. It is problematic that SNOMED CT does not support negation very well, since this is an important reason to represent clinical information by a terminology. E.g. valproate is used for medical therapy of prophylaxis of migraine, bipolar disorder, and epilepsy [18]. But in SNOMED CT |Valproate (Product)| is not defined as a migraine prophylaxis agent by a subtype relationship. Figure 10 illustrates that |Valproate (Prod-

uct)| is only inherited from two supertype parents. All SNOMED CT concepts within the |Pharmaceutical/biologic product (Product)| subtype hierarchy are primitive since these SNOMED CT concepts do not have sufficient defining characteristics. Since cardiac medication was only specified by fine-grained it was not possible to map the DE "'18 Hjertemedicin"' to a subtype descendant of |Drug therapy (Procedure)|. A SNOMED CT concept named something like "'Cardiac drug therapy"' was needed for this purpose.

Some DEs are subtypes of others

From figure 3 and 9 it is observed that some DEs are subtypes of others. E.g. this is true for the DE |Nitroglycerin (Product)| which has a vasodilating effect. Therefore, |Nitroglycerin (Product)| is a subtype descendant of |Hypotensive agent (Product)|. A query for "hypotensive" returns all information about antihypertensive medication - also medical therapy by use of nitroglycerin. This result may not be the intention of these two DEs.The DE "15) Nitroglycerin" should has been a data value of "17) Antihypertensiva". A similar pattern is found within the group of diagnoses. From figure 3 it is ascertained that |Cerebrovascular accident (Clinical finding)| was a subtype descendant of |Disorder of nervous system (Clinical finding)|. These examples show that some of the DEs may not be defined appropriately.

SNOMED CT provides optional level of granularity

SNOMED CT facilitates recording and representation of information with optional level of granularity. E.g. for the DEs of medication (see figure 9) |Hypotensive agent (Product)|, |Cardiac drug (Product)|, |Diuretic (Product)|, and |Nitroglycerin (Product)| are all inhereted from |Cardiovascular drug (Product)|. Therefore, the query "Cardiovascular" would retrieve any drug which is a subtype descendant of |Cardiovascular drug (Product)|. Thus, it is possible to retrieve information based on an optional level of granularity. This feature provides further/less subdivision of DEs dependent on the purpose of investigation and what level of granularity is appropriate for the recorded (size of) population.

If a dataset is collected and it is ascertained that normal distribution cannot be obtained based on the predefined DEs/data values, SNOMED CT provides redefinition of the DEs/data values based on a lower level of granularity, thus less DEs/data values.

Or, if 1000 samples are recorded in a research study to get an overview of the correlation between the number of medical products and the risk of syncope episodes, this could be ascertained by use of coarse-grained SNOMED CT concepts. If further 5000 samples were recorded later, more fine-grained SNOMED CT concepts could be used to investigate if the risk of syncope episodes is larger for certain pharmaceutical drugs and if there is any correlation between drug dose and risk of syncope episodes. Therefore, *if* the sample size increases *and* the study protocol allows to record information with a high level of granularity, it becomes possible to conduct detailed analyses.

Further, it is always possible to merge more categories (SNOMED CT concepts) to define a residual DE (like "22) Andet_medicin"). It is not possible to obtain the details of a residual class. What if a correlation between the DE "22) Andet_medicin" and the risk of syncope episodes was found? From the current design of the DEs and data values it is not possible to investigate if any *other* types of drugs increases the risk of syncope episodes. If the other types of medication were specified and mapped to SNOMED CT, it would be possible to conduct more detailed investigations based on initial findings.

Problems with internationalisation

Both the |Substance (Top-level concept)|- and the |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchies includes chemical constituents which are not approved by the Danish Health and Medicines Authority and used in DK. For that reason the current setup allows for selection of these unapproved concepts may be selected for specification of medical therapy, but these concepts do not have an ATC-code in SKS and cannot be linked to the ATC-classification. This problem could be solved by appropriate configuration of the CIS, to make it impossible to select these pharmaceutical products in the user interface.

The multiple parent paradigm - details and problems

The multi-axial hierarchy of SNOMED CT depicts the physiological reality and in that way provides more details than a classification system.

Some types of medicine have more than a single therapeutic effect. Diuretics are both a

renal- and a cardiovascular drug, since diuretics are used to decrease oedema caused by chronic renal insufficiency and/or chronic heart insufficiency. As illustrated on figure 9 SNOMED CT depicts this reality of diuretics.

Since the SNOMED CT concept |Diuretics (Product)| is primitive it is not possible to determine if the diuretic drug is prescribed to treat congestive heart failure or decreased renal function and in that way facilitate statistics by use of the terminology of SNOMED CT. For some statistical purposes, the primitive concepts may be problematic. E.g. conduction of a statistical distribution based on the queries "cardiovascular drug" and "renal drug" would retrieve the patients who receive diuretic therapy twice, thus induce redundancy.

In the ATC-classification of SKS, diuretics are defined as cardiovascular drugs, only. In a mono-axial classification, like ATC, it is not possible to let a class be inherited from more than a single parent. Thus, the ATC-classification only provides statistics based on the predefined dichotomy, which does not reflects all details of the real world.

Some antiepileptic drugs also have multiple therapeutic effects. E.g. therapeutic effects on bipolar disorder, neuropathic pain, anxiety, or prophylaxis of migraine. [18] Figure 10 shows an example of the antiepileptic drug named valproate which also has a therapeutic effect on bipolar disorder.

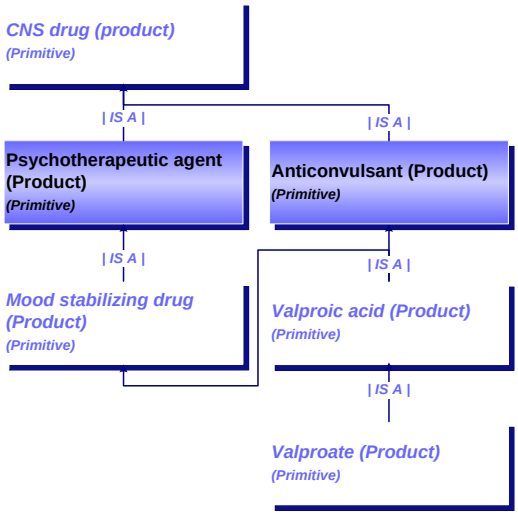


Figure 10. *Example of valproate which plays multiple therapeutic roles.*

In research these therapeutic agents which both have a therapeutic effect of epilepsy and psychiatric disorders could have influence of the results, since a patient treated for epilepsy could simultaneously be "treated for psychiatric disorders", unless it is unintented. The dataset provided for this project included the two separate DEs for antiepileptic medication and psychoactive drugs. These DEs do not necessarily reflect the reality of the effects of medical therapy.

Need of powerful mapping tools

To map the content of the dataset to SNOMED CT, it was necessary to get an overview of the subtype descendants, possible relationships according to the SNOMED CT Concept Model, and defining characteristics of candidate concepts. None of the software tools used in this project did support this need very well. These tools may not be intended to have these functionalities, because they are developed for other purposes.

HealthTerm does neither support post-coordination by qualification nor post-coordination by combination. Only post-coordination by refinement is supported. Further, HealthTerm does not has any functionality for graphical overviews of relationships between SNOMED CT concepts.

CliniClue has functionality for graphical overviews of relationships between SNOMED CT concepts. But these overviews does only show relationships between SNOMED CT concepts selected by the user. Thus, the user needs to know the allowed relationships on forehand and select post-coordinations based these relationships in order to create a graphical overview of these relationships. Thus, the graphical overviews of CliniClue were not very useful for the scope of this project. Further, the graphical overviews represents the SNOMED CT concepts and relationships in a way which is not optimal to provide a graphical overview.

For that reason it was necessary to achieve a thorough knowledge of SNOMED CT and to conduct drawings on paper of the defining relations between SNOMED CT concepts, to get an overview of pros and cons of selecting one set of SNOMED CT concepts prior to another.

Bugs in both the HealthTerm browser and CliniClue were found. In the HealthTerm browser some FSN-based queries did not retrieve any results, despite the concepts exist in SNOMED CT (and in the HealthTerm browser) and equal queries in CliniClue returned the results. E.g.

that was true for the queries; "rufinamide" and "lacosamide".

In the expression transformer of CliniClue it was possible to create non-sense post-coordinations which are not legal according to the concept model of SNOMED CT. E.g. it was possible to create the expression: |Syncope (Clinical finding)| |Has dose form (Linkage concept)| |Oral (Qualifier value)|.

4.1.2 Mapping was complicated and time-consuming

Mapping the research dataset was time-consuming. Especially mapping of the Medication group was time-consuming, since it was not obvious which subtype hierarchy to select from. Mapping guidelines which are well documented for mapping of research data would probably reduce the number of iterations during mapping. Further, a higher maturity level of the (|Pharmaceutical/Biologic product (Top-level concept|) subtype hierarchies may reduce the resources spend on mapping. Mapping was also complicated by the lack of powerful mapping tools which provides graphical overviews of subset clusters of SNOMED CT concepts and corresponding defining relationships. Therefore, it was necessary to draw subset clusters on paper to get an overview of possible defining relationships.

New revisions of SNOMED CT has low impact of subset clusters

The negative impact of new versions of a classification system is well known for data within the Danish National Patient Register (DNPR). The use of subset clusters to perform information retrieval is less sensitive to upgrades of SNOMED CT compared to a classification where each DE is defined by manually hardcoded specification of data values selected from different hierarchies. E.g. if a new anticonvulsant, pharmaceutical product is added to SNOMED CT this new product is automatically retrieved from the query "anticonvulsant drug". If the pharmaceutical drugs of a DE were manually specified, the new pharmaceutical drug must be added to the specification before the new drug can be retrieved from a query. Adding more defining relationships may not have influence as well. Only changes of the structure of the subtype hierarchies may have impact of information retrieval which is defined based on the structure of an older version of SNOMED CT.

4.1.3 Possibilities and limitations of selection of |Disorders (Top-level concepts)| for mapping of diagnoses

SNOMED CT provides high expressiveness and detail compared to the ICD-classification

Compared to the ICD-10 classification system of SKS, SNOMED CT provides an excellent subset cluster of syncope disorders and subtypes of syncope. In ICD-10 each type of syncope is found in different hierarchies, thus not creating a subset cluster feasible for information retrieval and computer processing. Unfortunately |Syncope (Clinical finding)| do not include subtype descendants of syncope caused by cardiac failure. The dichotomisation of syncope within SNOMED CT is a causal collection of different types of syncope according to pathophysiological-, clinical-, and external causes of syncope. The optimal dichotomisation of syncope for the purpose in this project was a dichotomisation according to the underlying pathophysiological cause and according to characteristic clinical symptoms, with a clear distinction between these two dichotomisations.

SNOMED CT shows that some DEs were broadly defined

Due to the semantical variety of the data values of "5) Neurologisk" this DE was broadly defined. Redefinition of "5) Neurologisk" might be appropriate to create a smaller subset cluster, since only disorders which may have an impact of syncope episodes were of interest.

4.1.4 Possibilities and limitations of selection of |Pharmaceutical/ products (Top-level concept)| by use of SNOMED CT

The OE of medication was mapped to a top-level concept. It was not considered to be highly problematic that the |Pharmaceutical/biologic product (Top-level concept)| was selected for the OE, since the SNOMED CT concepts selected for the DEs were fine-grained, thus created smaller subset clusters. Data values could *only* be selected from the subset clusters of mapped DEs and the smaller subset clusters delimited the range of options to select from. Further, since the DEs and the data values were mapped to a fine-grained SNOMED CT concept the mapped data was still represented with at least the same level of granularity.

Linkage to current standards

If the pharmaceutical products used are specified, it is possible to specify dose form and dose. The

chemical constituent is always provided, when medication is specified with at least the product category. Dose form is provided for VTMs and higher granularity, and strength is provided implicitly by the naming convention of VMPs and AMPs. It could be possible to implement linkage between the ATC-classification and SNOMED CT on VTM level.

4.2 Mapping of medication

Need of specification of the therapeutic role of medication

Non of the dichotomisations found in the |Pharmaceutical/biologic product (Top-level concept)|-, |Substance (Top-level concept)|- or the |Procedure (Top-level concept)| subtype hierarchies did exactly support the needs in this project. There may be two reasons for that. First, some of the DEs and the data values may not be defined appropriately. Second, the optimal dichotomisation for this project was a higher degree of dichotomisation according to the therapeutic role. In SNOMED CT it was not possible to specify the therapeutic role with a certain level of granularity. E.g. some types of anticonvulsant drugs may be used to treat all types of epilepsy, while others are only used to treat a given subtype of epilepsy. Further, some types of antiepileptic medication are also used for psychotherapeutic medication. Therefore it would be of interest to specify if a patient was treated with a given type of anticonvulsant drug to treat generalised epilepsy, focal epilepsy or a psychiatric disorder. But SNOMED CT does only provides specification of “anticonvulsant”.

In future the dichotomisation of the |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchy will be improved. Instead of subtype relationships between the groups of therapeutic products and the fine-grained SNOMED CT concepts, the groups of therapeutic products will be applied in a separate subtype hierarchy within the |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchy. Relationships between the pharmaceutical products and the therapeutic role which the pharmaceutical products plays will be specified by defining relationships. The new defining relationships are defined by the relationship; |Pharmaceutical/biologic product (Top-level concept)| |Plays therapeutic role (Linkage concept)| |Therapeuric product group (Product)|. This new defining attribute will replace some of the subtype relationships which are currently present. Further, the problems with dosages which currently cannot be determined from the defining relationship of a VPM will be solved. SNOMED CT will, to some extent, support numerical values to solve this problem. [17] Thus, the problems with primitive SNOMED CT concepts from the product hierarchy experienced in this project may be solved in future and SNOMED CT concepts from the |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchy may become fully defined - or at least - less primitive.

4.2.1 Selection of the top-level hierarchy for mapping of medication

For the international version of SNOMED CT, IHTSDO recommends mapping medication to the |Substance (Top-level concept)| subtype hierarchy [17]. The |Substance (Top-level concept)| subtype hierarchy does not provide representation of combinatorial medical preparations by a single SNOMED CT concept. Further, the |Substance (Top-level concept)| subtype hierarchy does not include pre-coordinated SNOMED CT concepts which specify dose and dose form of medication as |Pharmaceutical/Biologic product (Top-level concept)| does. Therefore, representing medication by use of the |Substance (Top-level concept)| subtype hierarchy provides less precise expressivity and is not as convenient as use of the |Pharmaceutical/Biologic product (Top-level concept)| subtype hierarchy.

Expression of medication by use of concepts from the |Procedure (Top-level concept)| subtype hierarchy supports post-coordination by use of subtype descendants of |Context values for actions (Qualifier value)|. But these concepts are defined for specification of the procedural status and these do not provide expressivity for context of medication. E.g. the subtype descendants of |Context values for actions (Qualifier value)| are "‘|To be done (Qualifier value)|"’ or "‘|Did not attend (Qualifier value)|"’. Therefore, it was not possible to post-coordinate the data values to the context values of procedures since these concepts are developed to support a procedural status. Further, representing medication by use of concepts from the hierarchy of |Procedure (Top-level concept)| causes pleonasm, since the DE should be represented by a post-coordination, since these concepts are similar to the concepts within the subtype hierarchy of |Pharmaceutical/biologic product (Top-level concept)|, with the only difference that these are procedures and not medical products.

The complications of mapping the DEs of the medication group show that the |Pharmaceutical/Biologic product (Top-level concept)| hierarchy may not have reached a high level of maturity yet.

4.2.2 Fully defined SNOMED CT concepts within the |Pharmaceutical/biologic product (Top-level concept)| hierarchy are needed

It is a problem that the intermediate SNOMED CT concepts within the |Pharmaceutical/biologic product (Top-level concept)| which have more than a single parent are primitive. It is not possible to refine these concepts to specify for what therapeutic purpose medication is prescribed. Therefore, recorded information cannot be uniquely assigned and this makes it impossible to conduct statistics which includes these SNOMED CT concepts. However, in the future these problems will be solved by already planned improvements of the dichotomisation of the |Pharmaceutical/biologic product (Top-level concept)| subtype hierarchy and the introduction of the |Plays therapeutic role (Linkage concept)| attribute [7].

Medical therapy may always be used for a reason. Either to treat a clinical finding (a disorder, side effects, or as preventive treatment) or as part of a procedure. This information may already be recorded for a patient. A relationship between a pharmaceutical product and this already recorded information may provide further details to represent medication. E.g. in the research dataset provided for this project most DEs within the diagnoses groups reflected the DEs within the group of medication. But if a patient has ischeamic heart disease and takes cardiac medication, it is not known for sure if the patient is treated with a cardiac drug due to ischaemic heart disease or another disorder. It becomes further complicated if the patient has multiple diagnoses, takes multiple pharmaceutical products, or if the pharmaceutical products have several therapeutic effects. Further investigation of how this feature may be supported within SNOMED CT is needed.

4.2.3 Problems of representing negation

SNOMED CT does not support representation of the context of medication. Particularly, it was experienced that it is problematic that SNOMED CT does not support negation of medication.

Therefore, the data values of the research dataset could not be mapped to SNOMED CT and instead the data values should be represented by an information model. This way of representing the data values causes difficulties of information retrieval. It it not possible to retrieve the patients who *have* and those who *do not have* a given type of disorder or takes a given type of medication, by use of queries which consist of SNOMED CT concepts, only. Further, it is well known that extraction of negation from free text (e.g. by NLP-techniques) is problematic since negation may be represented in several different ways, which causes errors of interpretation. Therefore, it is important to represent whether a patient *takes medication* or *not*, explicitly and in a standardised way.

SNOMED CT includes the |Finding context value (Qualifier value)| to represent the context of SNOMED CT concepts from the |Clinical finding (Top-level concept)| subtype hierarchy and the |Context values for actions (Qualifier value)| to represent the context of SNOMED CT concepts from the |Procedure (Top-level concept)| subtype hierarchy. Therefore, one solution for this problem could be the implementation of "'context values for medication"'. These context values should be generally applicable for all types of medication and -different contextual settings, and should support negation. The following list is a suggestion of how context values for medication could be defined:

- Unknown.
- Is on medication.
  - Routine use of medication. (Regular use every day)
  - When needed.
- Is not on medication.
  - Contraindicated.
    - Cave.
    - Due to side effects.
    - Due to other disorder.
    - Due to drug.
  - Not indicated.
  - Refused.
    - Due to side effect.
  - Insufficient therapeutic effect.
  - Subject of record does not follow the prescribed medical therapy plan.

This suggestion is formulated based on the results of this project and by inspiration of subtype descendants of |Finding context value (Qualifier value)|. It is important to note, that this suggestion of how to represent the context of medication is not evaluated or verified. Other concepts for

"'context values for medication"' and another dichotomisation could be defined. Further investigation of how the context of medication should be represented to support negation optimally, is needed.

However, it is important that new modifications of SNOMED CT are generally valid and does not change over time. Further, it is important that these modifications are considered as generally valid on an international level and are not influenced by differences in culture or procedural standards across countries. Else, routine configuration- and updates are required to keep the terminology up to date and national extensions (which are proprietary) may be needed.

4.3 Discussion of the methodology used in this project

4.3.1 Applicability of [1]s' mapping guideline for mapping mapping research data

Mapping the research dataset required instructions how to map medication and negations which is out of the scope of [1]s' mapping guideline. Therefore, it was not possible to use [1]s' mapping guideline for mapping these types of information from the research dataset. For mapping the research dataset it was logical to map according to the semantic type before mapping according to the data type, which is in reverse order of [1]s' mapping guideline. There are two reasons for this. First, the groups were primarily defined according to semantics and according to the quality criteria all DEs of a group should be mapped to the same subtype hierarchy. Second, for medication it was not obvious which subtype hierarchy to select from. The semantically mapping showed the possibilities and limitations of how to represent the DE. Therefore, it was not relevant to decide how to represent the DE before the subtype hierarchy was selected.

In this project the OEs were mapped to the LCP of the group. This provided a precise definition (and representation) of each group, because the OEs defined the semantics of the DEs and the OEs were as fine-grained as possible.

4.3.2 Discussion of grouping

Benefits of grouping

Grouping of the DEs ensured consistent mapping and improved the efficiency of mapping, since all DEs within a group may be mapped similarly. The OEs of a group was used to identify a LCP of a group. Grouping may make it easier to apply future changes or to add new DEs to the dataset. Further, grouping may reduce the risk that these actions lead to inconsistency. E.g. when adding a new DE the only challenge is to determine which group the new DE should be added to. The selected group defines how a new DE should be defined and mapped according to its data type.

Challenges of grouping

It was challenging and trivial to identify semantic groups where different data types appeared. E.g. "'table tilt test"' could be divided further to ensure the same data types within the group. It was not possible to create groups, where all groups fulfilled all the grouping criteria. Therefore, semantically grouping was preferred prior to grouping according to data types. All groups fulfilled not to create post-coordination by combination.

4.3.3 Discussion of the mapping procedure

The mapping process used in this project was specifically designed to support the needs in this project. Other mapping process designs than used in this project may be used for other mapping approaches.

4.3.4 Discussion of the quality criteria

The quality criteria provided a checklist for the expert review to evaluate the quality of mapping. The use of quality criteria may not be necessary if the mapping is performed by use of a mapping guideline which is well documented within the domain of mapping, since it may not be necessary to conduct a *meticulous* review of the mapped dataset if the mapping guideline is well documented to ensure consistent mapping.

4.4 Improvements of the methodology

4.4.1 Contextual knowledge of the dataset

For 1 DE ("'14) Medicin"') and all data values coded by numbers it was not possible to interpret their semantic meaning. Providing knowledge of the context of the research dataset may improve the mapping of the dataset, thus improve the possibilities of the mapped dataset. Contextual knowledge of the dataset could be obtained by access to the original study protocol, to know the purpose for which the research dataset was collected, and to discuss the meaning of the DEs of the dataset with the researchers who conducted

the clinical trial, which led to the recording of the dataset.

4.4.2 Redefinition of DEs

SNOMED CT highlights inappropriately defined DEs. From the results of this project it is showed that several DEs the dataset provided for this project were inappropriately defined. Redefinition of these DEs may improve the structure of the dataset and the possibilities for precise and efficient information retrieval.

4.4.3 Retrospective- vs. prospective mapping

Mapping the research dataset was conducted retrospectively. Therefore it was attempted to preserve the original DEs and -data values, to make it possible to map the recorded data to SNOMED CT. If the study protocol was prospectively designed based on SNOMED CT concepts, more benefits could be obtained. Further, with a proper prospective design problems with limitations and incompleteness of SNOMED CT could be avoided.

5. Conclusion

This project provides a methodology for grouping unstructured DEs, a set of quality criteria to review the mapping quality, and a mapping procedure to ensure consistent mapping. An example of how research data should be mapped to SNOMED CT and how to avoid the pitfalls of mapping research data to SNOMED CT are provided. Further, limitations of SNOMED CT are illustrated and clarified. This project demonstrates the outcomes of a SNOMED CT based research dataset.

5.1 Using [1]s' mapping guideline for mapping research data

The following elements of [1]s' mapping guideline were used in this project:

- Mapping of clinical result fields.
- The instructions of how to represent different data types by SNOMED CT concepts.
- The criteria of subtype relationship between the OE and the DEs.
- The idea of an iterative mapping process which was refined by a set of quality criteria.

To increase the applicability of [1]s' mapping guideline for mapping of research data (and other domains), it would make sense to divide the guideline into two parts which provides instructions of mapping according to semantics- and data types, respectively. In that way it may be optional if mapping according to semantics is performed before mapping according to data types or vice versa. Further, their mapping guideline needs to specify how medication and context values (of medication and clinical findings) should be mapped to SNOMED CT.

OEs- and DEs of medication should be mapped to the concepts from the |Pharmaceutical/biologic product (Top-level concept)| hierarchy.

Grouping is both a valuable and important methodology to apply for a dataset with non-grouped DEs.

5.2 Conclusion - What may SNOMED CT add to research data?

This project shows that SNOMED CT provides optional level of granularity to facilitate high flexibility of representing information from a research dataset. The multi-axial hierarchies and defining relationships provides more details of the real world, which may influence the results of clinical research studies. These features cannot be obtained by a classification system. In this project it was difficult to interpret the semantic meaning of the DEs of the dataset. SNOMED CT provides explicit semantic meaning, by the absolute position of SNOMED CT concepts in the subtype hierarchies and by the defining characteristics of a SNOMED CT concept. Mapping research data to SNOMED CT provides standardisation of the DEs and highlights if DEs are defined inappropriately, thus improves the data quality. Since SNOMED CT is an international terminology, it provides standardisation on an international level. Further, SNOMED CT provides multilingualism for efficient translation of clinical information, e.g. to translate research data for international meta-analyses or international large scale studies. Therefore, SNOMED CT facilitates comparability of data- and interoperability across countries.

As conclusion, despite of limitations due to incompleteness, inconsistency, and lack of support to represent negation, SNOMED CT provides several benefits which are feasible in clinical research. If possible, study protocols should always be designed based on SNOMED CT prospectively to obtain the maximal benefits of using SNOMED CT.

SNOMED CT based research data is one step towards semantic interoperability and efficient data extraction from EHR-systems, thus one step towards efficient, high-quality translational research and improved outcome of the clinical care process.

Further improvements of SNOMED CT are needed. This project highlights the need of development of a model to support negation and a higher maturity level of the |Pharmaceutical / biologic product (Top-level concept)|. In this project the DEs and the data values were not redefined. Future studies should investigate how to redefine the DEs which are inappropriately defined. Further development and evaluation of mapping guidelines are needed to provide well documented mapping guidelines which are proofed to ensure efficient and consistent mapping of research data to SNOMED CT.

This project only shows the outcome of a single case. Further investigation and evaluation of how to represent research data (especially context values and negation) by SNOMED CT are needed. This investigation includes mapping the remaining groups of the dataset, development of an information model, and implementation of this information model in a database. This project shows the benefits of SNOMED CT by theoretical examples. Investigation of a SCT based research dataset implemented in a database provides proof of the practical benefits of a SNOMED CT based research dataset.

References

[1] AR Rasmussen and KR Gøeg. Snomed ct implementation: Mapping guidelines facilitating reuse of data. *Methods of Information in Medicine*, (6):p. 1–10, 2012.

[2] J. Ingenerf and W. Giere. Concept-oriented standardization and statistics-oriented classification: Continuing the classification versus nomenclature controversy. *Methods of Information in Medicine*, 37(4):ch. 6, p. 527–539, 1998.

[3] ESC, Task Force for the Diagnosis and Management of Syncope of the European Society of Cardiology (ESC), European Heart Rhythm Association (EHRA), Heart Failure Association (HFA), Heart Rhythm Society (HRS), Moya A, Sutton R, Ammirati F, Blanc JJ, Brignole M, Dahm JB, Deharo JC, Gajek J, Gjesdal K, Krahn A, Massin M, Pepi M, Pezawas T, Ruiz Granell R, Sarasin F, Ungar A, van Dijk JG, Walma EP, and Wieling W. Guidelines for the diagnosis and management of syncope (version 2009). *European Heart Journal*, 30:p. 2631–2671, 2009.

[4] Jesper Mehlsen and Anne-Birgitte Mehlsen. Udredning og behandling af reflekssynkoper. *Ugeskrift for læger*, 170(9):p. 718–723, 2008.

[5] Jyotishman Pathak, Janey Wang, Sudha Kashyap, Melissa Basford, Rongling Li, Daniel R Masys, and Christopher G Chute. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the emerge network experience. *Journal of the American Medical Informatics Association*, 18, 2011.

[6] Dennis Lee, Nicolette de Keizer, Francis Lau, and Ronald Cornet. Literature review of snomed ct use. *Journal of the American Medical Informatics Association*, 0, 2013.

[7] International Health Terminology Standards Development Organisation IHTSDO. Snomed ct© technical implementation guide - january 2013 international release (gb english), 2013.

[8] Dennis Lee, Francis Lau, and Hue Quan. A method for encoding clinical datasets with snomed ct. *BMC Medical Informatics and Decision Making*, 10(53):12 pages, 2010.

[9] Dennis Lee, Ronald Cornet, Francis Lau, and Nicolette de Keizer. A survey of snomed ct implementations. *Journal of Biomedical Informatics*, 46, 2012.

[10] Geraldine Wade and S Trent Rosenbloom. Experiences mapping a legacy interface terminology to snomed ct. *BMC Medical Informatics and Decision Making*, 8(3):6 pages, 2008.

[11] International Health Terminology Standards Development Organisation IHTSDO. Snomed ct© user guide - january 2013 international release (gb english). User guide, 2013.

[12] Anna Vikström, Ylva Skånér, Lars-Erik Strender, and Gunnar H Nilsson. Mapping the categories of the swedish primary health care version of icd-10 to snomed ct concepts: Rule development and intercoder reliability

in a mapping trial. *BMC Medical Informatics and Decision Making*, 7, 2007.

[13] RH Dolin, KA Spackman, and D Markwell. Selective retrieval of pre-and post-coordinated snomed concepts. *American Medical Informatics Association*, pages p. 210–214, 2002.

[14] Ronald Cornet and Nicolette de Keizer. Forty years of snomed: a literature review. *BMC Medical Informatics and Decision Making*, 8, 2008.

[15] WHO Collaborating Centre for Drug Statistics Methodology. Antiepileptika. webpage, 2013.

[16] Statens Serum Institut. Det fælles medicinkort snitfladebeskrivelse version 1.4.0.4 2013-05-14. Technical manual, 2013.

[17] International Health Terminology Standards Development Organisation IHTSDO. Snomed ct© editorial guide - january 2013 international release (gb english). User guide, 2013.

[18] Per Sidenius. Antiepileptika. webpage, 2012.

Appendix: Figures in A3

Figure 3 and 6 in A3 format for clarification is attached on the following pages.